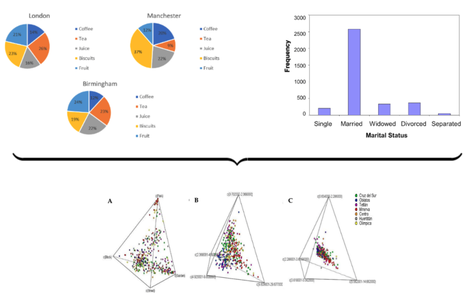
The Continuous Categorical: A novel simplex-valued exponential family

Elliott Gordon-Rodriguez, Gabriel Loaiza-Ganem, and John Cunningham

Department of Statistics, Columbia University; Layer6 AI

Motivation: compositional data



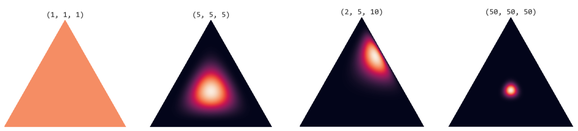
Simplex: $\mathbb{S}^K := \{\mathrm{x} \in \mathbb{R}_+^K : \sum_{i=1}^K x_i = 1\}$

Background: shortcomings of the Dirichlet

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

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

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

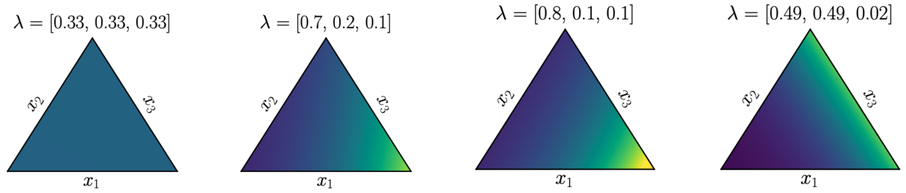


Solution: the continuous categorical

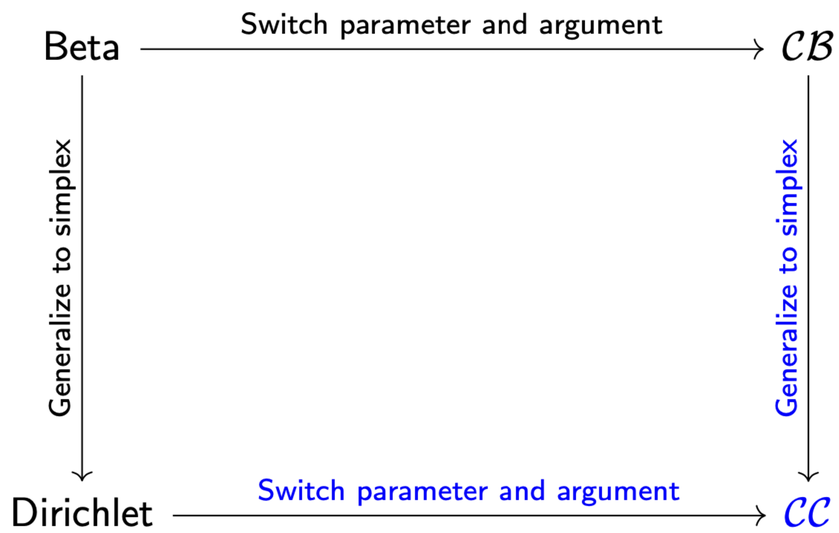
Definition: $\mathrm{x} \in \mathbb{S}^K$ follows a *continuous categorical* ($\mathcal{CC}$) distribution with parameter $\boldsymbol{\lambda} \in \mathbb{S}^K$ if:

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

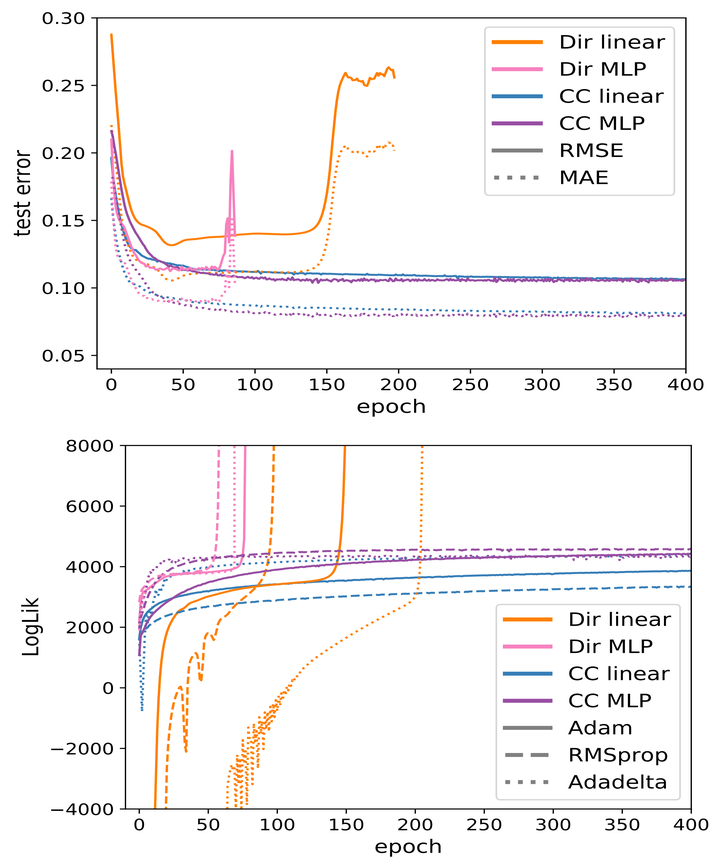
- **Extrema.** $\log p(\mathrm{x}; \boldsymbol{\lambda})$ is finite at the extrema of the simplex.
  $\therefore$ log-likelihood is well-defined in the presence of zeros.
- **Bias.** Re-write the $\mathcal{CC}$ density in canonical form
  $p(\mathrm{x}; \boldsymbol{\lambda}) \propto \exp\left(\sum_{i=1}^K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



Related distributions



Experiments: election data



Experiments: knowledge distillation