On a novel probability distribution for zero-laden compositional data

E. Gordon-Rodriguez[1,*], **G. Loaiza-Ganem**[2], **and J.P. Cunningham**[1]
[1]Columbia University, New York, USA; **eg2912@columbia.edu*
[2]Layer6 AI, Toronto, Canada

Summary

Recent advances in probabilistic machine learning have given rise to a new family of distributions on the simplex. This distribution, termed the *continuous categorical*, bears similarities with the Dirichlet in that it defines an exponential family with a particularly simple closed-form density. However, unlike the Dirichlet (or any log-ratio based method), the continuous categorical log-likelihood function is well defined even in the presence of zero-valued components, making this distribution a valid likelihood model for zero-laden compositional data without requiring imputation of the zeros. In this abstract, we review the key properties of our novel distribution and we present an application where it can be used for dimensionality reduction of compositional data. We also highlight some underexplored connections between the fields of machine learning and compositional data analysis, to which our novel distribution is closely related.

Key words: Simplex, probability distribution, exponential family, deep learning.

1 Introduction

While seldom referred to as such, Compositional Data (CoDa) appears in several machine learning tasks. For example, in knowledge distillation, a *student network* is trained to predict the softmax outputs of a *teacher network* (Hinton et al., 2015). Similarly, in *actor-mimic* reinforcement learning, a *mimic network* is trained to predict the policies of several *expert networks* across tasks (Parisotto et al., 2015). These examples amount to a compositional regression problem, where the prediction target is a simplex-valued object. The categorical cross-entropy is typically used as a loss function in such models. However, this loss function does not correspond to the log-likelihood of any probability model over the simplex, an observation that has given rise to the novel *continuous categorical* (CC) distribution (Gordon-Rodriguez et al., 2020).

Despite only having been discovered recently, the CC is arguably the simplest nontrivial distribution supported on the simplex. As such, we believe this distribution is of structural importance to the field of CoDa, and our primary goal is therefore to disseminate the CC among the CoDa community. We shall now present a brief overview of the CC and its theoretical properties, together with an application to illustrate how it can be used for learning low-dimensional representations of compositional data. Note that other applications including knowledge distillation and actor-mimic learning have been explored in Gordon-Rodriguez et al. (2020). However, a limiting factor in such applications is that the normalizing constant of the CC distribution, despite its mathematical simplicity, suffers from unstable floating-point behavior in high dimensions. These numerical properties have been explored in more depth by Gordon-Rodriguez et al. (2022). However, new theoretical or computational advances are still needed to enable high dimensional applications, for example microbiome studies.

2 The continuous categorical

The CC distribution is defined by the density:

$$f(\mathbf{x};\boldsymbol{\lambda}) = \frac{1}{C(\boldsymbol{\lambda})}\prod_{j=1}^{K}\lambda_j^{x_j}, \tag{1}$$

where both the random variable and the parameter live in the simplex, i.e., $\mathbf{x},\boldsymbol{\lambda}\in\Delta^{K-1}=\{\mathbf{x}\in\mathbb{R}_+^K : \sum_j x_j = 1\}$. Notice the similarity between the CC density and that of the Dirichlet; we essentially "switch" the role of the parameter and the data. While the Dirichlet density either diverges or vanishes whenever $\mathbf{x}$ contains zero elements, the CC density remains finite and positive, meaning it can be used as a likelihood model for zero-laden CoDa, without having to apply zero-replacement methods such as Martín-Fernández et al. (2000). The term $C(\boldsymbol{\lambda})$ is the normalizing constant, which was shown to be equal to (Gordon-Rodriguez et al., 2020):

$$C(\boldsymbol{\lambda}) = \sum_{k=1}^{K}\frac{\lambda_k}{\prod_{j\neq k}\log\frac{\lambda_k}{\lambda_j}}, \tag{2}$$

provided $\lambda_j \neq \lambda_k$ for all $j \neq k$. Unlike the Dirichlet, this normalizing constant depends on elementary functions only. However, despite its mathematical simplicity, $C(\boldsymbol{\lambda})$ can be hard to compute numerically in high dimensions, due to positive and negative summands resulting in catastrophic cancellation (Gordon-Rodriguez et al., 2022). Indeed, deriving a numerically stable algorithm for evaluating Equation 2 in high dimensions remains an open problem.

Note that, when taken as a probability mass function over the vertices of the simplex (i.e., one-hot vectors), the CC density (Eq. 3) becomes the categorical distribution and the normalizing constant vanishes. Since the log-likelihood of the categorical distribution is identical to the (negative) cross-entropy loss for categorical data (i.e., classification), similarly the log-likelihood of the CC distribution can be thought of as a probabilistic equivalent of the cross-entropy loss for simplex-valued data (i.e., compositional regression):

$$\log f(\mathbf{x};\boldsymbol{\lambda}) = -\log C(\boldsymbol{\lambda}) + \sum_{j=1}^{K} x_j \log\lambda_j. \tag{3}$$

The implication for probabilistic machine learning is as follows: *changing the sample space from one-hot vectors to the simplex results in an additional log-normalizer term in the loss function.*

The CC also defines an exponential family of probability distributions, where the natural parameter corresponds to the log-ratio $\eta_j = \log\lambda_j/\lambda_K$, which is unconstrained real-valued. In this parameterization, our density takes a particularly concise form:

$$f(\mathbf{x};\boldsymbol{\eta}) = \frac{1}{C(\boldsymbol{\eta})}e^{\boldsymbol{\eta}^\top\mathbf{x}}. \tag{4}$$

Thus, the CC can also be viewed as a multivariate exponential distribution restricted to the simplex. By the standard theory of exponential families, it follows that the moments of our distribution can be computed by differentiating the normalizing constant, which in practice we implement using automatic differentiation. The CC also admits efficient sampling algorithms, based on combining independent draws from the 1-dimensional case (Gordon-Rodriguez et al., 2020).

3 Related work

The use of the cross-entropy loss to model non-categorical, simplex-valued data is widespread in machine learning, including knowledge distillation (Hinton et al., 2015), actor-mimic reinforcement learning (Parisotto et al., 2015), generative models of images (Kingma and Welling, 2013) and transfer learning (Tzeng et al., 2015). However, it is mainly recent works that have questioned the underlying assumptions and, in turn, motivated the CC distribution (Loaiza-Ganem and Cunningham, 2019; Gordon-Rodriguez et al., 2020; Wong et al., 2021). Many other recent works in the CoDa field have also been inspired by advances in machine learning (Tolosana-Delgado et al., 2019; Quinn et al., 2020; Gordon-Rodriguez et al., 2022; Templ, 2021). Moreover, we believe the converse is highly promising as well. As an example, deep generative models with discrete latent variables are typically optimized by means of the Gumbel-softmax distribution on the simplex (Jang et al., 2016). However, it was later demonstrated that the logistic-normal distribution of Aitchison and Shen (1980) provided a better model (Potapczynski et al., 2020).

4 Application

The CC has shown good empirical properties in regression models with compositional outputs. For example, Gordon-Rodriguez et al. (2020) obtain improved distillation performance by using a CC log-likelihood, relative to other loss functions on the simplex. We now present an orthogonal application, where we demonstrate the use of the CC for dimensionality reduction. For simplicity, and leveraging the fact that the CC defines an exponential family, we focus on exponential family PCA (Collins et al., 2001). Nevertheless, our approach can be extended straightforwardly to more general factor models, including Bayesian priors (Tipping and Bishop, 1999), nonlinear representations (Kingma and Welling, 2013), and so forth.

Recall standard PCA aims to find a subspace that minimizes the total squared distance from the data points $\mathbf{x}_i$ to their projections $\boldsymbol{\theta}_i$ in the subspace, i.e.:

$$\min \sum_{i=1}^{n} \|\mathbf{x}_i - \boldsymbol{\theta}_i\|^2. \tag{5}$$

In the case of exponential family PCA, the data $\mathbf{x}_i$ need not be unconstrained real-valued, but rather are assumed to be drawn from an exponential family. As such, it is inappropriate to minimize Euclidean distance; we instead maximize the log-likelihood of the appropriate exponential family. The projections $\boldsymbol{\theta}_i$ are still constrained to a low-dimensional linear subspace, but they are mapped to the parameters of the exponential family by a (possibly nonlinear) *link function*. In this regard, $\boldsymbol{\theta}_i$ plays the same role as the linear predictor in a generalized linear model (McCullagh and Nelder, 2019). In practice, we parameterize $\boldsymbol{\theta}_i$ using a subspace basis $\mathbf{v}_1, \ldots, \mathbf{v}_D$, together with the per-datapoint representations $\mathbf{z}_1, \ldots, \mathbf{z}_n$, so that $\boldsymbol{\theta}_i = \sum_{j=1}^{D} z_{ij}\mathbf{v}_j$. In the case of the CC family, we choose the *canonical* link function which sets the linear predictor to be equal to the natural parameter of the exponential family, i.e., $\boldsymbol{\eta}_i = \boldsymbol{\theta}_i$, following the notation of Equation 4. Thus, our objective function takes on a particularly simple form:

$$\max_{\{\mathbf{z},\mathbf{v}\}} \sum_{i=1}^{n} \log f\left(\mathbf{x}_i; \boldsymbol{\eta}_i\right) = \max_{\{\mathbf{z},\mathbf{v}\}} \sum_{i=1}^{n} \left\{ -\log C \left(\sum_{j=1}^{D} z_{ij}\mathbf{v}_j \right) + \sum_{j=1}^{D} z_{ij}\mathbf{v}_j^\top \mathbf{x}_i \right\}. \tag{6}$$

This model can be fitted straightforwardly using an automatic differentiation library such as Tensorflow (Abadi et al., 2016).[1]

[1] See `https://github.com/cunningham-lab/cb_and_cc/tree/master/cc/pca` for the details of our implementation.

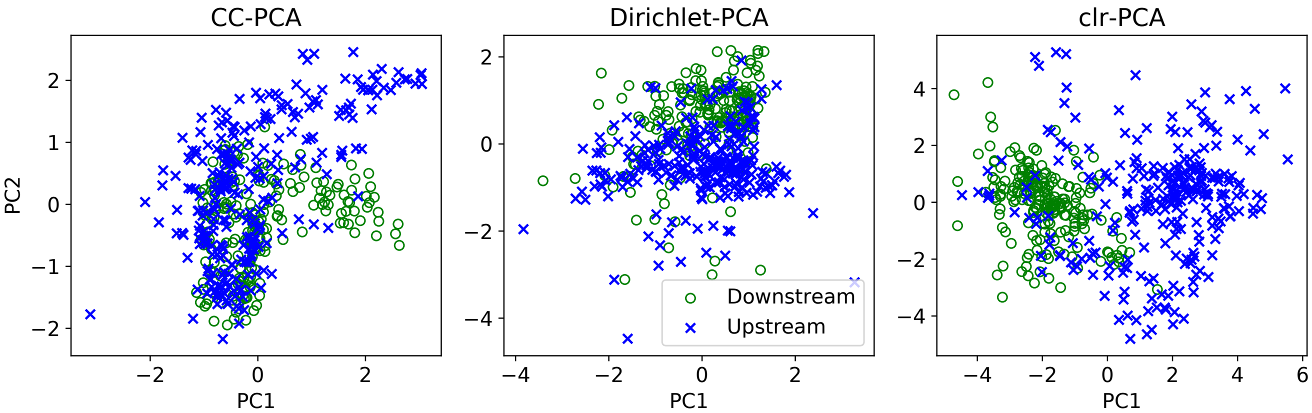

Figure 1: Learned principal components for the Llobregat river hydrochemical dataset (Tolosana-Delgado et al., 2005). Each point denotes a hydrochemical sample, color-coded depending on whether the sample was taken upstream (closer to the source) or downstream (closer to the mouth).

As a point of comparison, we also consider exponential family PCA with a Dirichlet likelihood. In this case, our setup is the same except for the link function and objective. We now have that $\boldsymbol{\alpha}_i = e^{\boldsymbol{\theta}_i}$, where $\boldsymbol{\alpha}_i$ is the Dirichlet parameter for the ith observation. The objective function is now the Dirichlet log-likelihood, which can be written in terms of $\mathbf{z}$ and $\mathbf{v}$ as follows:

$$\max_{\{\mathbf{z},\mathbf{v}\}} \sum_{i=1}^{n} \left\{ -\log B\left(e^{\sum_{j=1}^{D} z_{ij}\mathbf{v}_j}\right) + \sum_{j=1}^{D} (e^{z_{ij}\mathbf{v}_j} - \mathbf{1})^\top \log \mathbf{x}_i \right\}, \tag{7}$$

where $B(\cdot)$ denotes the multivariate beta function, and the exponentiation and logarithm are taken componentwise. Notice that the logarithm in Equation 7 requires zero-imputation be applied prior to optimization, unlike the CC case.

In addition to (CC and Dirichlet) exponential family PCA, we consider standard PCA applied after transforming the data using the *centered log ratio* (clr) (Aitchison, 1982). The compare the 3 methods on 2 datasets. The first is a hydrochemical dataset from the Llobregat river basin, which has been studied by Tolosana-Delgado et al. (2005); Wang et al. (2008). The second is a multi-party election dataset from the UK 2019 general election (Uberoi et al., 2019). We plot the principal components of the two datasets in Figures 1 and 2, respectively. While all 3 methods are capable of recovering important latent structure in the data, the CC model is the only one that does not require zero-imputation upstream of the analysis. However, note that these datasets are low-dimensional ($K < 15$); the CC model cannot be applied on high-dimensional CoDa, such as microbiome data, due to the numerical instability of the normalizing constant Gordon-Rodriguez et al. (2022). We hope that further work on the relevant computational techniques will enable the application of the CC on hundred- or even thousand-dimensional problems.

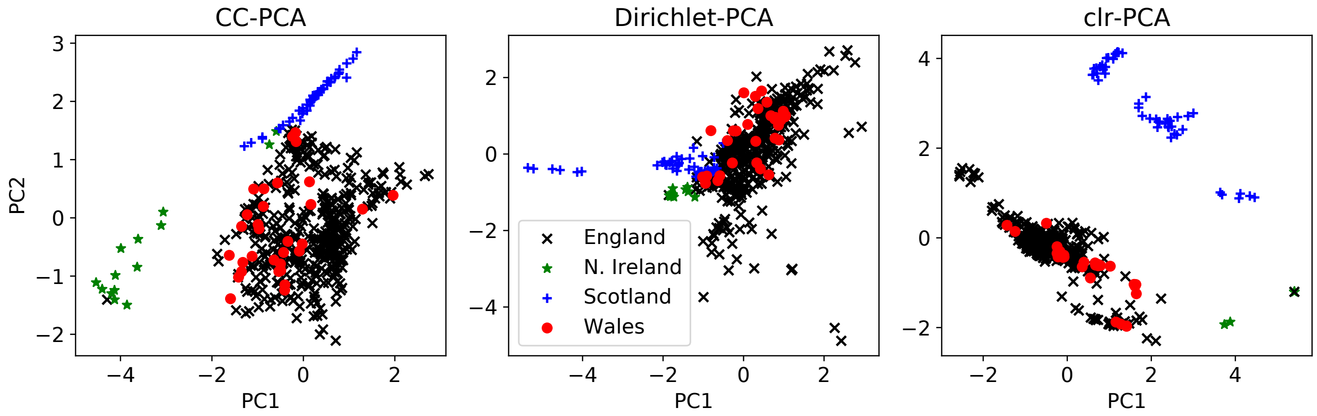

Figure 2: Learned principal components for the UK election data (Uberoi et al., 2019). Each point denotes an electoral constituency, color-coded by country.

References

Abadi, M., A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological) 44*(2), 139–160.

Aitchison, J. and S. M. Shen (1980). Logistic-normal distributions: Some properties and uses. *Biometrika 67*(2), 261–272.

Collins, M., S. Dasgupta, and R. E. Schapire (2001). A generalization of principal components analysis to the exponential family. *Advances in neural information processing systems 14*.

Gordon-Rodriguez, E., G. Loaiza-Ganem, and J. Cunningham (2020). The continuous categorical: a novel simplex-valued exponential family. In *International Conference on Machine Learning*, pp. 3637–3647. PMLR.

Gordon-Rodriguez, E., G. Loaiza-Ganem, G. Pleiss, and J. P. Cunningham (2020). Uses and abuses of the cross-entropy loss: Case studies in modern deep learning. In *Proceedings of "I Can't Believe It's Not Better!" at NeurIPS Workshops*, pp. 1–10.

Gordon-Rodriguez, E., G. Loaiza-Ganem, A. Potapczynski, and J. P. Cunningham (2022). On the normalizing constant of the continuous categorical distribution. *arXiv preprint arXiv:2204.13290*.

Gordon-Rodriguez, E., T. P. Quinn, and J. P. Cunningham (2022). Learning sparse log-ratios for high-throughput sequencing data. *Bioinformatics 38*(1), 157–163.

Hinton, G., O. Vinyals, J. Dean, et al. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531 2*(7).

Jang, E., S. Gu, and B. Poole (2016). Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Kingma, D. P. and M. Welling (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Loaiza-Ganem, G. and J. P. Cunningham (2019). The continuous bernoulli: fixing a pervasive error in variational autoencoders. In *Advances in Neural Information Processing Systems*, pp. 13266–13276.

Martín-Fernández, J., C. Barceló-Vidal, and V. Pawlowsky-Glahn (2000). Zero replacement in compositional data sets. In *Data analysis, classification, and related methods*, pp. 155–160. Springer.

McCullagh, P. and J. A. Nelder (2019). *Generalized linear models*. Routledge.

Parisotto, E., J. L. Ba, and R. Salakhutdinov (2015). Actor-mimic: Deep multitask and transfer reinforcement learning. *arXiv preprint arXiv:1511.06342*.

Potapczynski, A., G. Loaiza-Ganem, and J. P. Cunningham (2020). Invertible gaussian reparameterization: Revisiting the gumbel-softmax. *Advances in Neural Information Processing Systems 33*, 12311–12321.

Quinn, T., D. Nguyen, S. Rana, S. Gupta, and S. Venkatesh (2020). Deepcoda: personalized interpretability for compositional health data. In *International Conference on Machine Learning*, pp. 7877–7886. PMLR.

Templ, M. (2021). Artificial neural networks to impute rounded zeros in compositional data. In *Advances in Compositional Data Analysis*, pp. 163–187. Springer.

Tipping, M. E. and C. M. Bishop (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61*(3), 611–622.

Tolosana-Delgado, R., N. Otero, V. Pawlowsky-Glahn, and A. Soler (2005). Latent compositional factors in the llobregat river basin (spain) hydrogeochemistry. *Mathematical Geology 37*(7), 681–702.

Tolosana-Delgado, R., H. Talebi, M. Khodadadzadeh, and K. Van den Boogaart (2019). On machine learning algorithms and compositional data. In *Proceedings of the 8th International Workshop on Compositional Data Analysis, Terrassa, Spain*, pp. 3–8.

Tzeng, E., J. Hoffman, T. Darrell, and K. Saenko (2015). Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE international conference on computer vision*, pp. 4068–4076.

Uberoi, E., C. Baker, and R. Cracknell (2019). General election 2019: Full results and analysis. *Parliament UK. July*.

Wang, H.-Y., Q. Yang, H. Qin, and H. Zha (2008). Dirichlet component analysis: feature extraction for compositional data. In *Proceedings of the 25th international conference on Machine learning*, pp. 1128–1135.

Wong, J. H., I. Abramovski, X. Xiao, and Y. Gong (2021). Diarisation using location tracking with agglomerative clustering. *arXiv preprint arXiv:2109.10598*.