# A New Distribution on the Simplex with Auto-Encoding Applications

Columbia University

New York Genome Center

Andrew Stirn, Tony Jebara, David A Knowles
{andrew.stirn,jebara,daknowles}@cs.columbia.edu

## Contributions

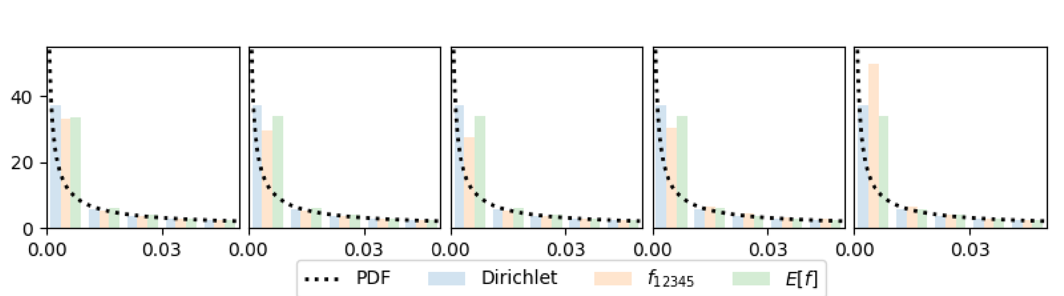
We develop a surrogate distribution for the Dirichlet that offers explicit, tractable reparameterization, the ability to capture sparsity, and has barycentric symmetry (exchangeability) properties equivalent to the Dirichlet.

## Stick Breaking Process

**Algorithm 1** Ordered Stick-Breaking

**Require:** $K \geq 2$
**Require:** base dist. $p_i(v; a_i, b_i) \; \forall \, i \in [K]$
**Require:** ordering (permutation) $o$
  Sample: $v_{o_1} \sim p_{o_1}(v; a_{o_1}, b_{o_1})$
  Assign: $x_{o_1} \leftarrow v_{o_1}, i \leftarrow 2$
  **while** $i < K$ **do**
    Sample: $v_{o_i} \sim p_{o_i}(v; a_{o_i}, b_{o_i})$
    Assign: $x_{o_i} \leftarrow v_{o_i}\left(1 - \sum_{j=1}^{i-1} x_{o_j}\right)$
    Assign: $i \leftarrow i + 1$
  **end while**
  Assign: $x_{o_K} \leftarrow 1 - \sum_{j=1}^{K-1} x_{o_j}$
  **return** $x$

## Kumarswamy Stick Breaks




Bias for a 5-dimensional sparsity-inducing Dirichlet approximation using $\alpha = \frac{1}{5}(1,1,1,1,1)$. We maintain histograms for each sample dimension for three methods: Dirichlet, Kumaraswamy stick-breaks with a fixed order, Kumaraswamy stick-breaks with a random ordering. Note the bias on the last dimension when using a fixed order. Randomizing order eliminates this bias.

## An Exchangeable Dirichlet Surrogate

Let $f_o(x_{o_1}, \ldots, x_{o_K}; \alpha_{o_1}, \ldots, \alpha_{o_K})$ be the joint density of $K$ random variables returned from algorithm 1 with $p_i(v; a_i, b_i) \equiv \text{Kumaraswamy}(x; \alpha_i, \sum_{j=i+1}^{K} \alpha_j)$ and some ordering $o$, then our proposed distribution for the $(K-1)$-simplex is

$$\text{MV-Kumaraswamy}(x; \alpha) = \mathop{\mathbb{E}}_{o \sim \text{Uniform}(O)}[f_o(x_{o_1}, \ldots, x_{o_K}; \alpha_{o_1}, \ldots, \alpha_{o_K})]$$

**Corollary 1** *Let $S \subseteq \{1, \ldots, K\}$ be the set of indices $i$ where for $i \neq j$ we have $\alpha_i = \alpha_j$. Define $A = \{1, \ldots, K\} \setminus S$. Then, $\mathbb{E}_{o \sim \text{Uniform}(O)}[f_o(x_{o_1}, \ldots, x_{o_K}; \alpha_{o_1}, \ldots, \alpha_{o_K})]$ is symmetric across barycentric axes $x_a \; \forall \, a \in A$ (i.e. it is exchangeable).*

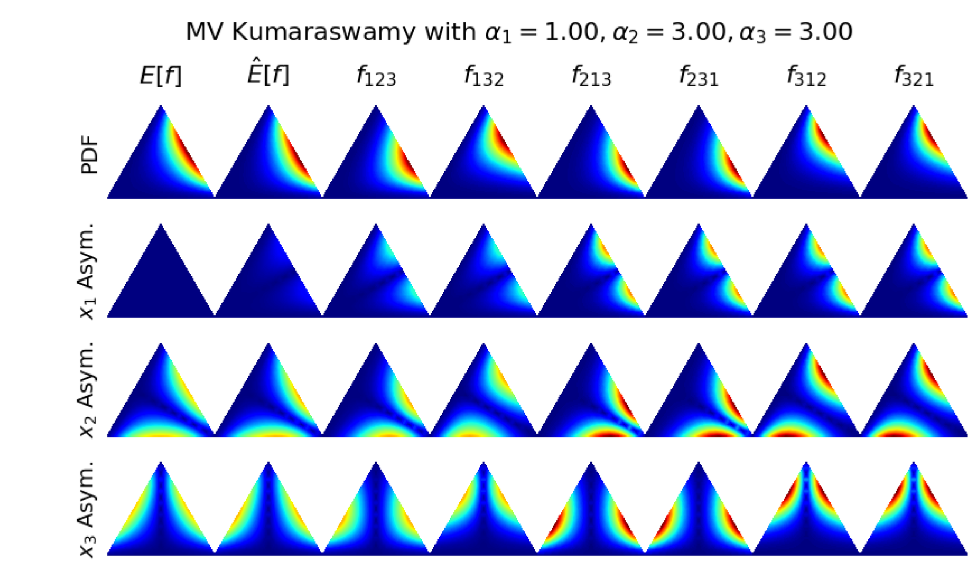
## Semi-supervised Variational Auto-encoding Tasks

We specify the a generative process with partially observed labels $y$. We we fit this model with an VAE. Each method varies in its treatment of the variational posterior $q(\pi; \alpha_\phi(x))$

$$\pi \sim \text{Dirichlet}(\pi; \alpha), \qquad z \sim \mathcal{N}(z; 0, I),$$
$$y|\pi \sim \text{Discrete}(y; \pi), \qquad x|y, z \sim p(x|f_\theta(y, z)),$$

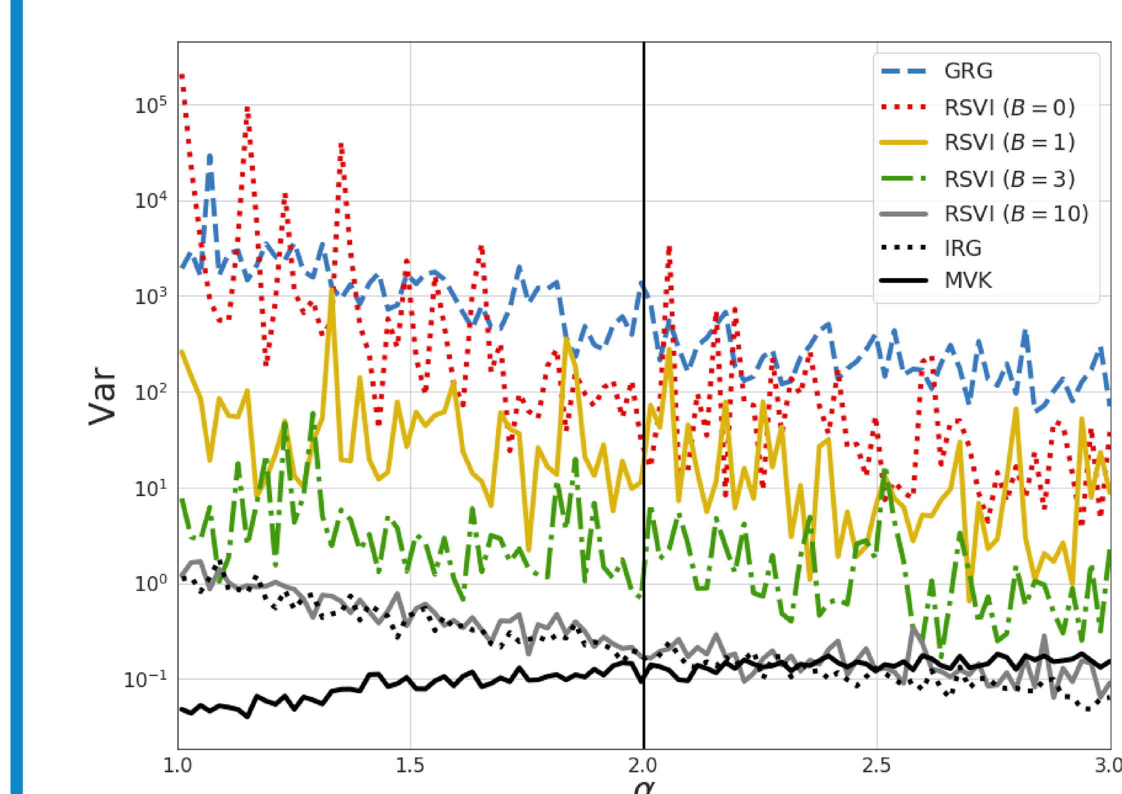
| Experiment | Method | Error | $p$-value | Log Likelihood | $p$-value |
|---|---|---|---|---|---|
| MNIST 10 trials 600 labels $\dim(z) = 0$ | MV-Kum. | $0.099 \pm 0.011$ | $--$ | $-6.4 \pm 6.3$ | $--$ |
| | IRG[1] | $0.097 \pm 0.008$ | 0.72 | $-7.8 \pm 7.1$ | 0.64 |
| | Kumar-SB[2] | $0.248 \pm 0.009$ | $1.05 \times 10^{-17}$ | $-6.5 \pm 6.3$ | 0.95 |
| | Softmax | $0.093 \pm 0.009$ | 0.24 | $-6.5 \pm 6.2$ | 0.95 |
| MNIST 10 trials 600 labels $\dim(z) = 2$ | MV-Kum. | $0.043 \pm 0.005$ | $--$ | $45.06 \pm 0.92$ | $--$ |
| | IRG[1] | $0.044 \pm 0.006$ | 0.89 | $45.69 \pm 0.38$ | 0.06 |
| | M2 (ours) | $0.098 \pm 0.014$ | $5.37 \times 10^{-10}$ | Not collected | $--$ |
| | Kumar-SB[2] | $0.138 \pm 0.015$ | $1.65 \times 10^{-13}$ | $44.33 \pm 1.65$ | 0.24 |
| | Softmax | $0.042 \pm 0.003$ | 0.40 | $45.14 \pm 0.73$ | 0.82 |
| MNIST 10 trials 600 labels $\dim(z) = 50$ | MV-Kum. | $0.018 \pm 0.004$ | $--$ | $116.58 \pm 0.68$ | $--$ |
| | IRG[1] | $0.018 \pm 0.004$ | 0.98 | $116.57 \pm 0.43$ | 0.97 |
| | M2 (ours) | $0.020 \pm 0.003$ | 0.32 | Not collected | $--$ |
| | Kumar-SB[2] | $0.071 \pm 0.008$ | $2.58 \times 10^{-13}$ | $116.22 \pm 0.33$ | 0.15 |
| | Softmax | $0.018 \pm 0.003$ | 0.87 | $116.24 \pm 0.45$ | 0.21 |
| | M2$^\dagger$[3] | $0.049 \pm 0.001$ | $--$ | Not reported | $--$ |
| | M1 + M2$^\dagger$[3] | $0.026 \pm 0.005$ | $--$ | Not reported | $--$ |
| SVHN 4 trials 1000 labels $\dim(z) = 50$ | MV-Kum. | $0.288 \pm 0.025$ | $--$ | $669.69 \pm 0.37$ | $--$ |
| | IRG[1] | $0.291 \pm 0.017$ | 0.85 | $668.93 \pm 0.53$ | 0.06 |
| | M2 (ours) | $0.396 \pm 0.010$ | $1.86 \times 10^{-04}$ | Not collected | $--$ |
| | Kumar-SB[2] | $0.707 \pm 0.012$ | $8.10 \times 10^{-08}$ | $669.03 \pm 0.43$ | 0.06 |
| | Softmax | $0.332 \pm 0.009$ | 0.02 | $669.55 \pm 0.11$ | 0.49 |
| | M1 + M2$^\dagger$[3] | $0.360 \pm 0.001$ | $--$ | Not reported | $--$ |

## Dirichlet Approximation




2-simplex with Kumaraswamy sticks

## Gradient Variance




Variance of the ELBO's gradient's first dimension for Categorical data with 100 dimensions and a Dirichlet prior. Others fit a Dirichlet. We fit a MV-Kumaraswamy using $K = 100$ samples (linear complexity) from Uniform($O$) to Monte-Carlo approximate the full expectation.

## References & Code

Paper and references available at:
`arxiv.org/abs/1905.12052`
Source code available at:
`github.com/astirn/MV-Kumaraswamy`

# References

[1] M. Figurnov, S. Mohamed, and A. Mnih, "Implicit reparameterization gradients," in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 441–452. [Online]. Available: http://papers.nips.cc/paper/7326-implicit-reparameterization-gradients.pdf

[2] E. Nalisnick and P. Smyth, "Stick-breaking variational autoencoders," *International Conference on Learning Representations (ICLR)*, Apr 2017. [Online]. Available: http://par.nsf.gov/biblio/10039928

[3] D. P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., 2014, pp. 3581–3589. [Online]. Available: http://papers.nips.cc/paper/5352-semi-supervised-learning-with-deep-generative-models.pdf