

# Efficient Parameter Estimation for Semi-Continuous Data: An Application to Independent Component Analysis

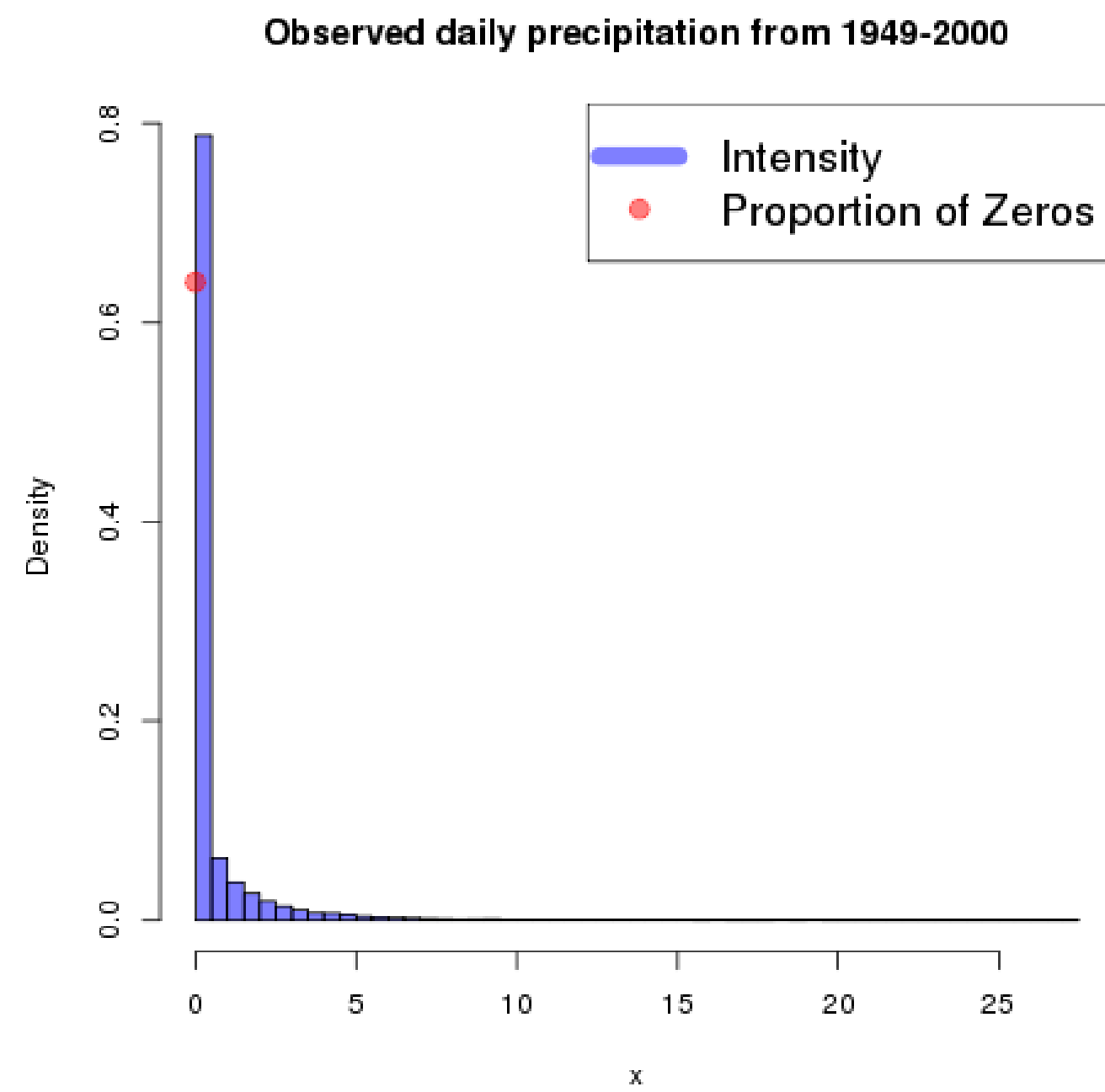
Sai K. Popuri<sup>1</sup>, Zois Boukouvalas<sup>2</sup>

<sup>1</sup>Walmart Labs, San Bruno, CA 94066

<sup>2</sup>American University, Department of Mathematics and Statistics, Washington, DC 20016

## Introduction

Semi-continuous data have a point mass at zero and are continuous with positive support. Such data arise naturally in several real-life situations like daily rainfall at a location, sales of durable goods among many others. Therefore, efficient estimation of the underlying probability density function (PDF) is of significant interest.



## Contribution

- ▶ We present an estimation method for semi-continuous data based on the maximum entropy principle.
- ▶ We demonstrate its successful application in developing a new Independent Component Analysis (ICA) algorithm, [ICA-Semi-continuous Entropy Maximization \(ICA-SCEM\)](#).
- ▶ We present a theoretical analysis of the proposed estimation technique and using simulated data we demonstrate the superior performance of ICA-SCEM over classical ICA algorithms.

## Estimation using entropy maximization

The PDF of a [semi-continuous random variable](#)  $Y$  can be written as<sup>1</sup>

$$p(y | \gamma, \theta) = \gamma\delta(y) + (1 - \gamma)\delta^*(y)g(y | \theta),$$

where  $g(y | \theta)$  is a PDF of a continuous random variable with support on  $(0, \infty)$ ,  $\gamma$  is the point mass at zero,  $\delta(y)$  is the indicator function, and  $\delta^*(y) = 1 - \delta(y)$ .

### Maximum Entropy Principle:

$$\max_{p(y)} H(p(y)) = - \int_I p(y) \log p(y) \mu(dy)$$

$$\text{s.t. } \int_I h_i(y)p(y) \mu(dy) = \alpha_i, \text{ for } i = 1, \dots, K,$$

where  $h_i(y)$  are measuring functions,

$\alpha_i = \sum_{t=1}^T h_i(t)/T$  are the sample averages, and  $K$  denotes the total number of measuring functions.

Using the maximum entropy principle estimate the distribution that maximizes the entropy of  $Y$ .

For known  $\gamma$ , the distribution that maximizes the entropy of  $Y$  is given by<sup>1</sup>

$$p(y) = \gamma\delta(y) + (1 - \gamma)\delta^*(y)g^*(y),$$

where  $g^*$  maximizes the entropy of a continuous random variable with support  $(0, \infty)$  subject to the constraints  $\int_0^\infty h_i(z)g(z)dz = \frac{\alpha_i}{1-\gamma}$ ,  $i = 1, \dots, K$ .

## Example

Suppose we have a sample of size  $n$  of semi-continuous data  $\{y_1, \dots, y_n\}$ ,  $\gamma$  is set to the proportion of zeroes in the data, and

$$\alpha_2 = \frac{1}{n} \sum_{i:y_i>0} y_i \text{ and } \alpha_3 = \frac{1}{n} \sum_{i:y_i>0} \log(y_i).$$

The resulting MaxEnt distribution is given by

$$f(y | \gamma, \kappa, \theta) = \gamma\delta(y) + (1 - \gamma)\delta^*(y) \frac{y^{\kappa-1} e^{-y/\theta}}{\theta^\kappa \Gamma(\kappa)},$$

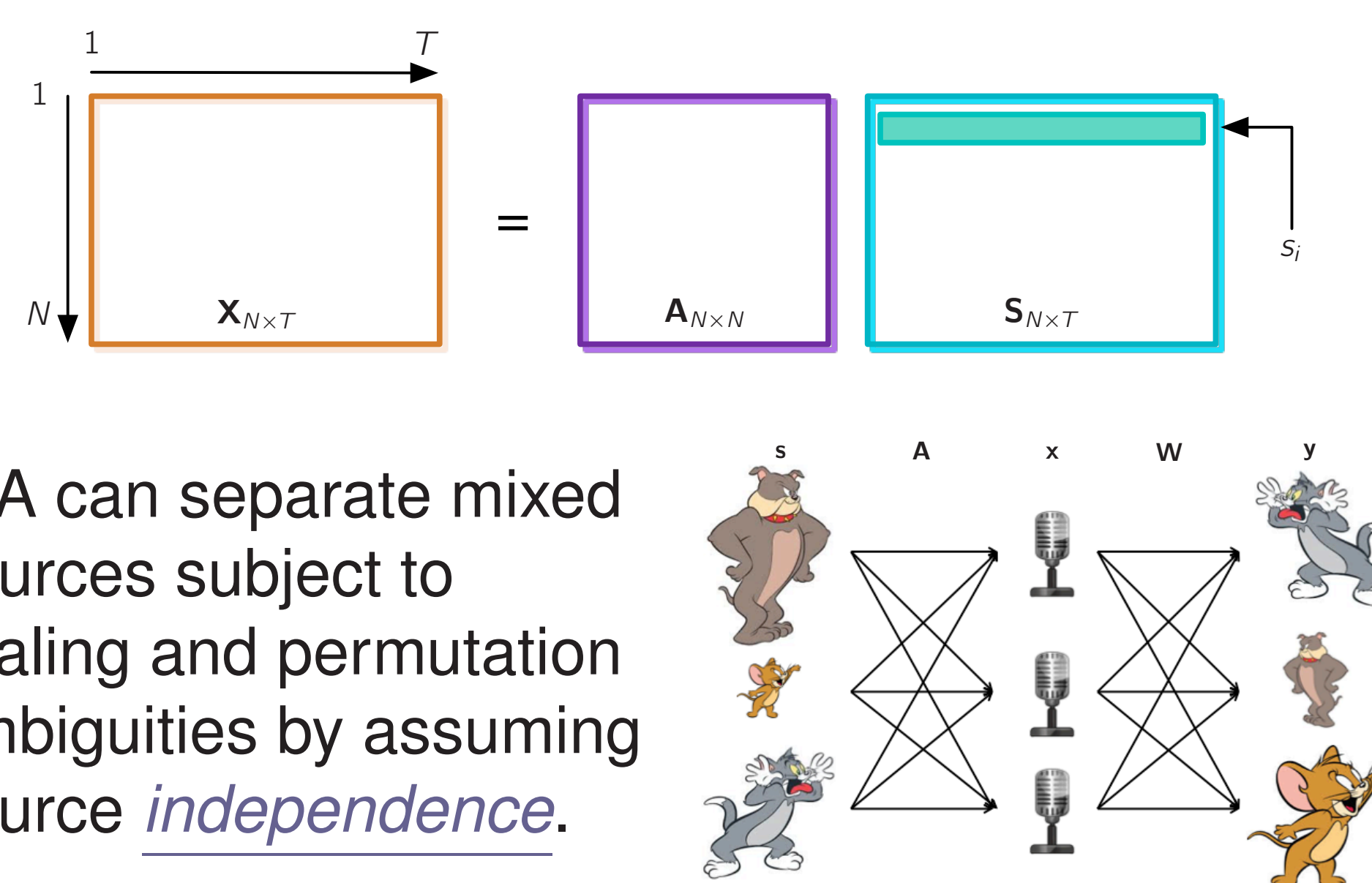
where  $\theta$  and  $\kappa$  are solutions to

$$\alpha_2 = \kappa\theta \text{ and } \alpha_3 = \psi(\kappa) + \log(\theta),$$

where  $\psi(\cdot)$  is the digamma function.

## Application to ICA

Generative model:  $\mathbf{x} = \mathbf{A}\mathbf{s}$ , where  $\mathbf{x}$  are the observations and  $\mathbf{s}$  are the latent sources linearly mixed by matrix  $\mathbf{A}$ .



ICA can separate mixed sources subject to scaling and permutation ambiguities by assuming source [independence](#).

In order to estimate  $\mathbf{W}$ , we [minimize](#) the mutual information (MI) among the source estimates<sup>2</sup>  $y_1, \dots, y_N$

$$J_{ICA}(\mathbf{W}) = \sum_{n=1}^N H(y_n) - \log |\det(\mathbf{W})| - H(\mathbf{x}),$$

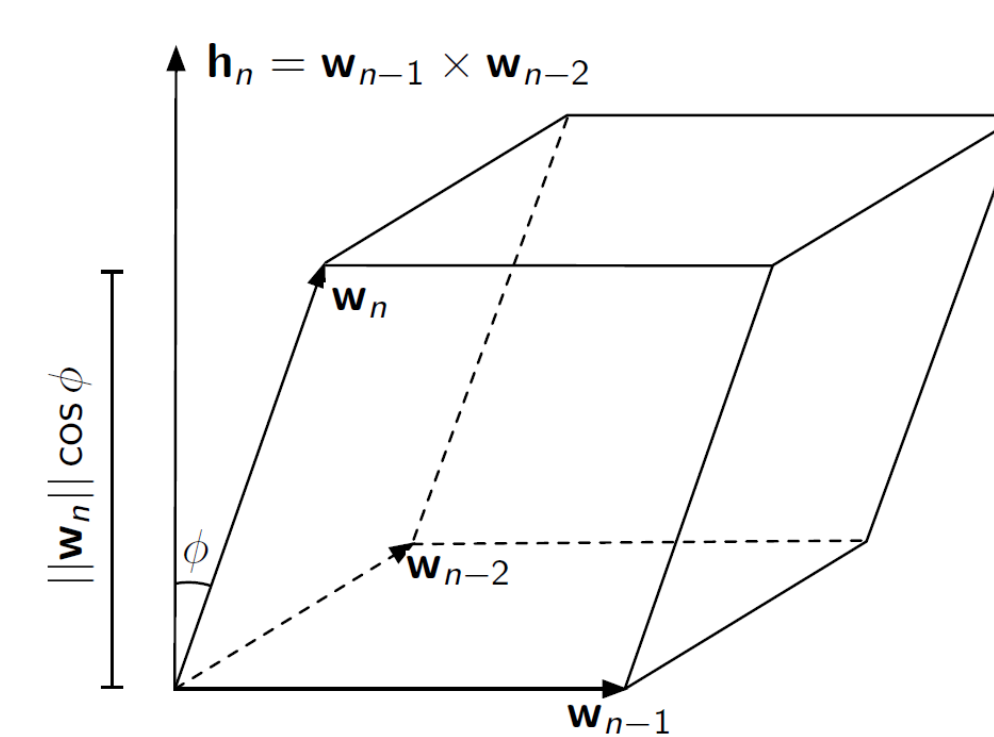
where  $H(y_n) = -E\{\log(p(y_n))\}$ .

Each  $y_n$  is assumed to be semi-continuous.

## Development of ICA-SCEM algorithm

Decoupling the MI cost function enables for the development of effective algorithms<sup>2</sup>.

This is achieved by expressing the volume of the parallelepiped,  $|\det(\mathbf{W})|$ , as the product of the area of its base and its height<sup>3</sup>.



The cost function with respect to each  $\mathbf{w}_n$  is given by

$$J_{ICA}(\mathbf{W}) = \sum_{n=1}^N H(y_n) - \log |(\mathbf{h}_n^\top \mathbf{w}_n)| - \log |\det(\mathbf{W}_n \mathbf{W}_n^\top)| - H(\mathbf{x}). \quad (1)$$

The gradient of (1) can be written in the decoupled form

$$\frac{\partial J(\mathbf{W})}{\partial \mathbf{w}_n} = -E\{\phi(y_n)\mathbf{x}\} - \frac{\mathbf{h}_n}{\mathbf{h}_n^\top \mathbf{w}_n}, \quad (2)$$

where  $\phi(y_n) = \frac{\partial \log p(y_n)}{\partial y_n}$ . As can be seen in (2), each gradient direction depends directly on the corresponding estimated source PDF and

$$\frac{\partial \log p(y_{n,t})}{\partial y_{n,t}} = \begin{cases} 0, & \text{if } y_{n,t} = 0 \\ \frac{\partial \log g(y_{n,t}|\theta_{n,t})}{\partial y_{n,t}}, & \text{if } y_{n,t} > 0. \end{cases}$$

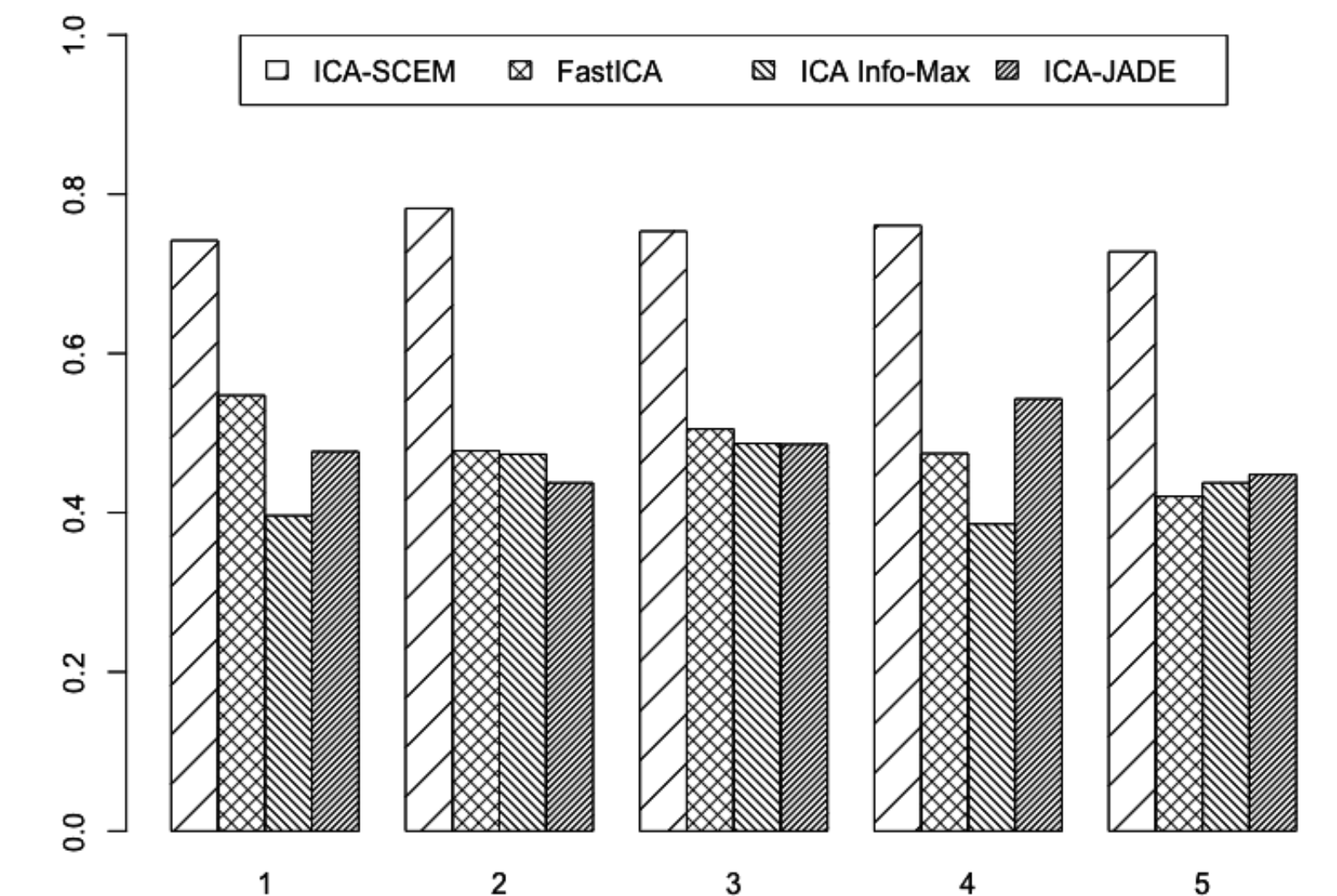
and  $\phi(\mathbf{y}_n) = [\frac{\partial \log p(y_{n,1})}{\partial y_{n,1}}, \dots, \frac{\partial \log p(y_{n,T})}{\partial y_{n,T}}]^\top$  is a vector of partial derivatives of dimension  $T$ .

## Experimental results

**Simulation 1:** Data for each source is generated using the two-part gamma distribution

$$f(y | \gamma, \kappa, \theta) = \gamma\delta(y) + (1 - \gamma)\delta^*(y) \frac{y^{\kappa-1} e^{-y/\theta}}{\theta^\kappa \Gamma(\kappa)},$$

where  $\gamma = 0.6$ ,  $\theta = 1$ , and  $\kappa = 1$

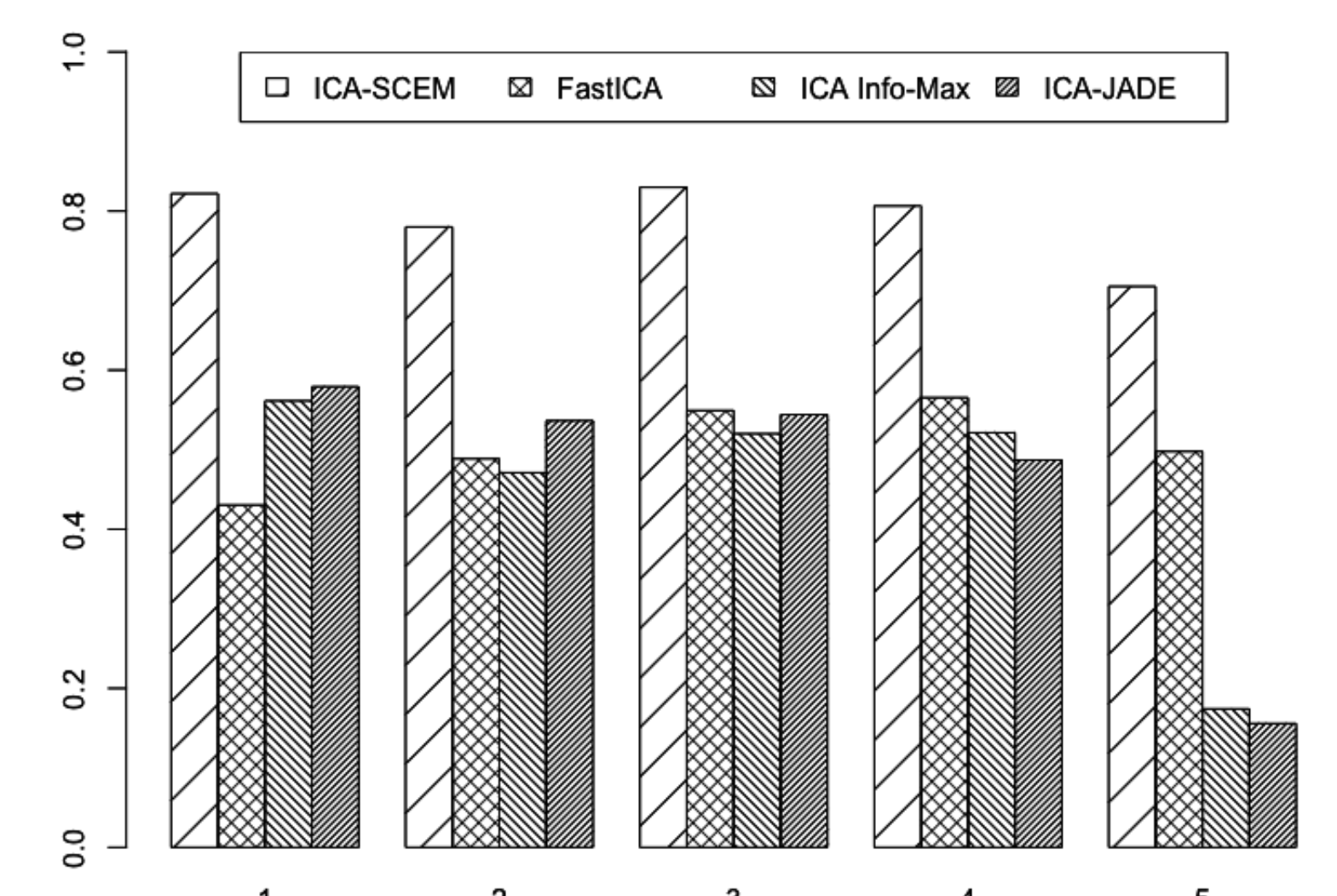


**Simulation 2:** Data for the first two out of five sources are generated using the two-part gamma model and for the rest of the three sources are generated using the following two-part lognormal distribution

$$f(y | \gamma, \mu, \sigma) = \gamma\delta(y) + (1 - \gamma)\delta^*(y) \frac{1}{y} \phi\left(\frac{\log(y) - \mu}{\sigma}\right),$$

where data for the five sources are generated according the following parameter choices

Source	$\gamma$	$\kappa$	$\theta$	$\mu$	$\sigma$
1	0.6	1	1	-	-
2	0.4	1	2	-	-
3	0.6	-	-	0	1
4	0.5	-	-	0.5	0.5
5	0.4	-	-	1	2



ICA-SCEM performs the best among well known ICA algorithms in terms of separation performance

## Conclusion and future directions

An efficient density estimation method for semi-continuous data was presented and a new ICA algorithm for semi-continuous data, ICA-SCEM, is proposed.

### Future Directions:

- ▶ Comparisons of ICA-SCEM, with ICA algorithms that exploit the sparsity of the data as well as non-negative source separation based methods.
- ▶ Multivariate extensions could be developed by considering multivariate distributions for the continuous part with element-wise Bernoulli probabilities determining the presence of zeros.

<sup>2</sup> T. Adali, M. Anderson, and G.-S. Fu, "Diversity in independent component and vector analyses: Identifiability, algorithms, and applications in medical imaging," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 18-33, May 2014.

<sup>3</sup> Z. Boukouvalas, Y. Levin-Schwartz, R. Mowakeaa, G.-S. Fu, and T. Adali, "Independent Component Analysis Using Semi-Parametric Density Estimation Via Entropy Maximization," *In 2018 IEEE Statistical Signal Processing Workshop (SSP)*, pp. 403-407, 2018.

<sup>1</sup> S. K. Popuri, "Prediction Methods for Semi-continuous Data with Applications in Climate Science," *Ph.D. thesis, University of Maryland, Baltimore County*, 2017.