# Multivariate Density Estimation Using Low-Rank Fejér-Riesz Factorization

**Paris A. Karakasis and Nicholas D. Sidiropoulos**

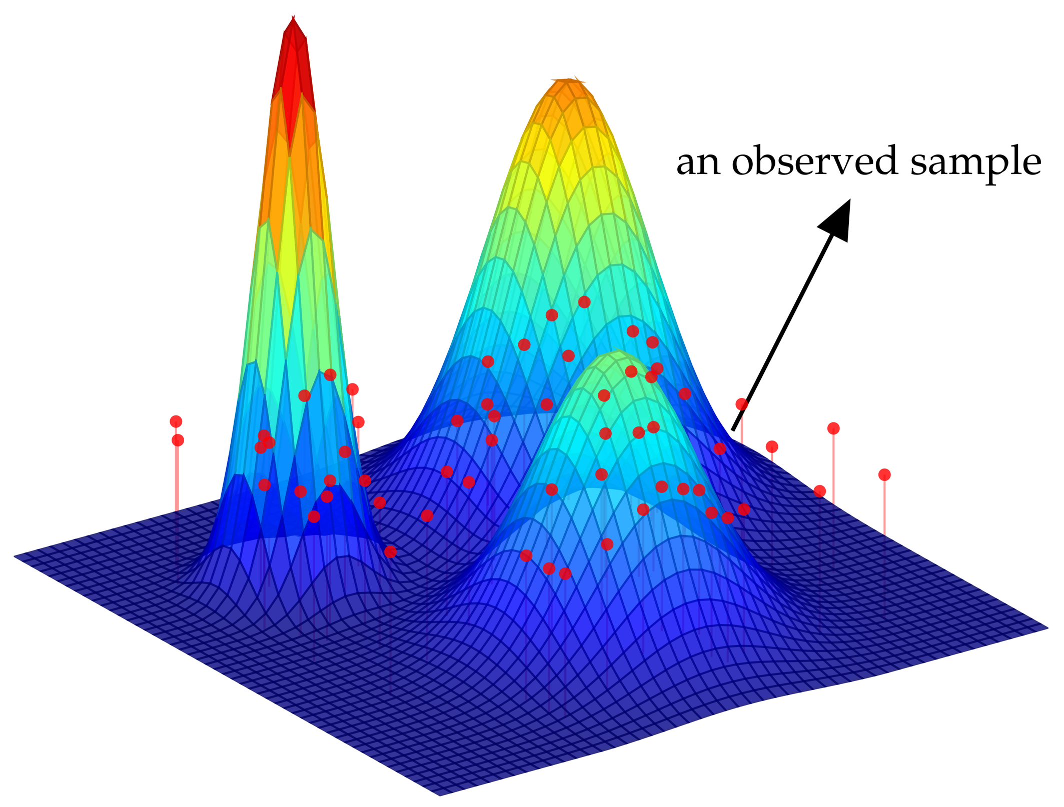
University of Virginia, Charlottesville, USA

## Introduction

**• Probability Density Estimation**

▷ Estimating a probability density from samples

▷ The most general & difficult problem in Statistical Learning!




**• Challenges & Popular Options**

⊗ Curse of Dimensionality **vs** Practical Sample Complexity

▷ **Parametric models**:
– Exponential Family Distributions, Bayesian Nets...

▷ **Nonparametric models**:
– KDE, Histograms, Parzen Windows, GMMs, OSDE...

▷ **Neural Network based density evaluation models**:
– RNADE, NF, MAF, Variational Autoencoders...

**• Orthogonal Series Density Estimation (OSDE)**

▷ **1D case**:

Learn $\widehat{f}_X : [0,1] \to \mathbb{R}$ using $\widehat{f}_X(x) = \sum_{k=-K}^{K} \widehat{\phi}_X[k] e^{j2\pi kx}$

▷ **ND case**:

Learn $\widehat{f}_{\mathbf{X}} : [0,1]^N \to \mathbb{R}$ using $\widehat{f}_{\mathbf{X}}(\mathbf{x}) = \sum_{k_1=-K}^{K} \cdots \sum_{k_N=-K}^{K} \widehat{\mathbf{\Phi}}_{\mathbf{X}}[\mathbf{k}] e^{j2\pi \mathbf{k}^T \mathbf{x}}$

$\widehat{\mathbf{\Phi}}_{\mathbf{X}}$ is a tensor → Canonical Polyadic Decomposition (CPD)!

▷ **Amiridi et al.**: Then, the following model emerges

$$\widehat{f}_{\mathbf{X}}(\mathbf{x}) = \sum_{r=1}^{R} \mathbf{P}_H[r] \prod_{n=1}^{N} \widehat{f}_{X_n|H=r}(x_n)$$

Easy Marginalization ✓ Cond. distr. & moments in closed-form ✓

Integration to one ✓ Nonnegativity ✗

## Proposed Approach

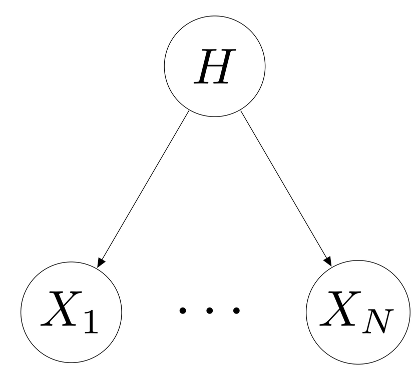
**• Proposed Model**

▷ Extending the Fejér-Riesz parametrization to the multivariate case

$$\widehat{f}_{\mathbf{X}}(\mathbf{x}) = \sum_{r=1}^{R} \widehat{\mathbf{P}}_H[r] \prod_{n=1}^{N} \widehat{f}_{X_n|H=r}(x_n),$$

- each $\widehat{f}_{X_n|H=r} : [0,1] \to \mathbb{R}_+$ admits $\widehat{f}_{X_n|H=r}(x) = \sum_{k=-K}^{K} \widehat{\phi}_{n,r}[k] e^{j2\pi kx}$
- $\widehat{\mathbf{P}}_H \in \mathbb{R}^R$: the PMF of a latent & discrete random variable $H$
- Naive Bayes interpretation:

Conditional independence of $\{X_n\}_{n=1}^N$ given $H$




▷ Ensuring that $\widehat{\mathbf{P}}_H$ is a proper PMF: $\widehat{\mathbf{P}}_H[r] = \frac{e^{\boldsymbol{\theta}_H[r]}}{\sum_{t=1}^{R} e^{\boldsymbol{\theta}_H[t]}}, \quad \forall r$

▷ Ensuring that each $\widehat{f}_{X_n|H=r}$ is a proper PDF

- Integration to one: $\int_0^1 \widehat{f}_{X_n|H=r}(x)\,dx = 1 \longrightarrow \widehat{\phi}_{n,r}[0] = 1$
- Nonnegativity: Fejér-Riesz Representation Theorem

$$\widehat{f}_{X_n|H=r}(x) \geq 0 \text{ iff } \exists\, \widehat{\boldsymbol{\psi}}_{n,r} \in \mathbb{C}^{K+1} \text{ s.t. } \widehat{f}_{X_n|H=r}(x) = \left| \sum_{k=0}^{K} \widehat{\boldsymbol{\psi}}_{n,r}[k] e^{j2\pi kx} \right|^2$$

- Combine them:

$$\widehat{\phi}_{n,r}[k] = \frac{\sum_{m=\max(k,0)}^{\min(K,K+k)} \widehat{\boldsymbol{\psi}}_{n,r}[m] \widehat{\boldsymbol{\psi}}^*_{n,r}[m-k]}{\left\| \widehat{\boldsymbol{\psi}}_{n,r} \right\|_2^2}$$

**• Learning the Model**

- **Constraint-free problem formulation for learning bona fide PDFs**
- **Compatible with stochastic optimization algorithms**
- **Scalable algorithm:** $\mathcal{O}(M_B RKN)$, where $M_B$ is the batch size
- **Ability to learn even from incomplete datasets**

▷ Given $M$ potentially incomplete samples & a mask $\mathbf{Q} \in \{0,1\}^{M \times N}$

▷ Minimize the empirical negative log-likelihood (NLL)

$$\min_{\{\widehat{\boldsymbol{\psi}}_{n,r}\}_{n=1,r=1}^{N,R}, \boldsymbol{\theta}_H} -\frac{1}{M} \sum_{m=1}^{M} \log \left( \sum_{r=1}^{R} \widehat{\mathbf{P}}_H[r] \prod_{n:\mathbf{Q}[m,n]=1}^{N} \widehat{f}_{X_n|H=r}(x_n^{(m)}) \right)$$

## Experimental Evaluation

**• Dataset**

- Synthetic dataset: $N = 10$, $M = 1000$, drawn from a mixture of $30$ Gaussian distributions with random diagonal covariance matrices

**• Baselines**

- Kernel Density Estimator (KDE)
- Gaussian Mixture Model (GMM)
- Real-valued Non-Volume Preserving Transformation based Density Estimator (RealNVP)[1]
- Masked Autoregressive Flow based Density Estimator (MAF) [2]
- CPD based Density Estimator (Amiridi et al.)[3]

| | NLL | MSE | MAE |
|---|---|---|---|
| KDE | -14.30± 0.04 | **0.0021 ± 0** | **0.0337 ±0** |
| GMM | -12.06± 1.89 | 0.0025 ± 0 | 0.0386 ±0 |
| MAF | -11.79± 0.60 | 0.0022 ± 0 | 0.0350 ± 0 |
| RealNVP | -10.93± 0.54 | 0.0026 ± 0 | 0.0386 ± 0 |
| Amiridi et al. | **-16.18± 0.07** | 0.0029± 0 | 0.0390 ± 0 |
| **Proposed** | **-14.68± 0.01** | **0.0020±0** | **0.0319±0** |

**Table 1:** Comparison in terms of NLL, MSE, and MAE

[1] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density Estimation using Real NVP," arXiv:1605.08803, 2016
[2] Papamakarios, T. Pavlakou, and I. Murray, "Masked Autoregressive Flow for Density Estimation," Advances in Neural Information Processing Systems, vol. 30, 2017
[3] M. Amiridi, N. Kargas, and N. D. Sidiropoulos, "Low-Rank Characteristic Tensor Density Estimation Part I: Foundations," IEEE Transactions on Signal Processing, vol. 70, pp. 2654–2668, 2022