MATHÉMATIQUES
VISION
APPRENTISSAGE

Stir it up! Mixture Density Networks

Antoine Debouchage [1] [2]   Valentin Denée [2]   Clement Wang [1] [2]

[1]CentraleSupélec   [2]Ecole Normale Supérieure Paris-Saclay

Introduction - Mixture Density Networks

Mixture Density Network (MDN) is the combined structure of a feed-forward network and a mixture model.

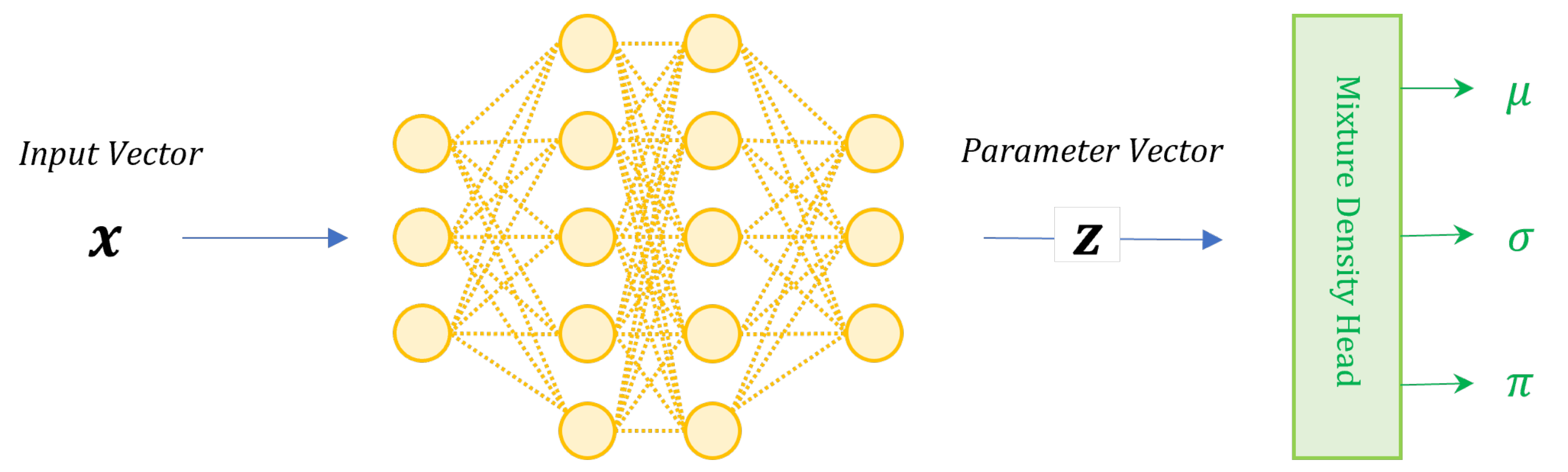


Figure 1. Architecture of the Mixture Density Network

$$p(t|x) = \sum_{i=1}^{m} \alpha_i(x)\phi_i(t|x) \qquad \phi_i(t|x) = \frac{1}{(2\pi)^{c/2}\sigma_i(x)^c} \exp\left(-\frac{||t-\mu_i(x)||^2}{2\sigma_i(x)^2}\right)$$

Instead of predicting a single point estimate for a target variable, MDNs predict a probability distribution characterized by a mixture of simpler distributions (such as Gaussians). The output of an MDN consists of parameters (mean μ, variance σ, and weights π) for these component distributions.

Mixture Density Network loss

$$\boxed{\mathcal{L}oss(y_{\text{true}}, \alpha, \mu, \sigma) = -\log\left[\sum_{i=1}^{m} \alpha_i \frac{1}{\sqrt{2\pi}^c \sigma_i^c} \exp\left(-\frac{\|y_{\text{true}} - \mu_i\|^2}{2\sigma_i^2}\right)\right]}$$

is directly derived from the **negative log-likelihood** of the predicted distribution.

MDN loss vs MSE loss

Standard Neural Networks approximate conditionnal average

$$E^S(w) = \frac{1}{2}\sum_{k=1}^{c} \iint (f_k(x;w) - t_k)^2 p(t,x)\,dt\,dx$$

$$f_k(x, w^*) = \langle t_k \mid x\rangle = \int t_k(t)p(t|x)\,dt$$

- Unlike MSE loss, which predicts a single point, MDNs can capture **multiple peaks or modes** in the data distribution, providing a more accurate representation of complex and diverse data patterns.
- MDNs explicitly model uncertainty by predicting a **probability distribution** rather than a single point estimate. This is particularly valuable when dealing with inherently uncertain data, allowing for a richer representation of uncertainty in predictions.

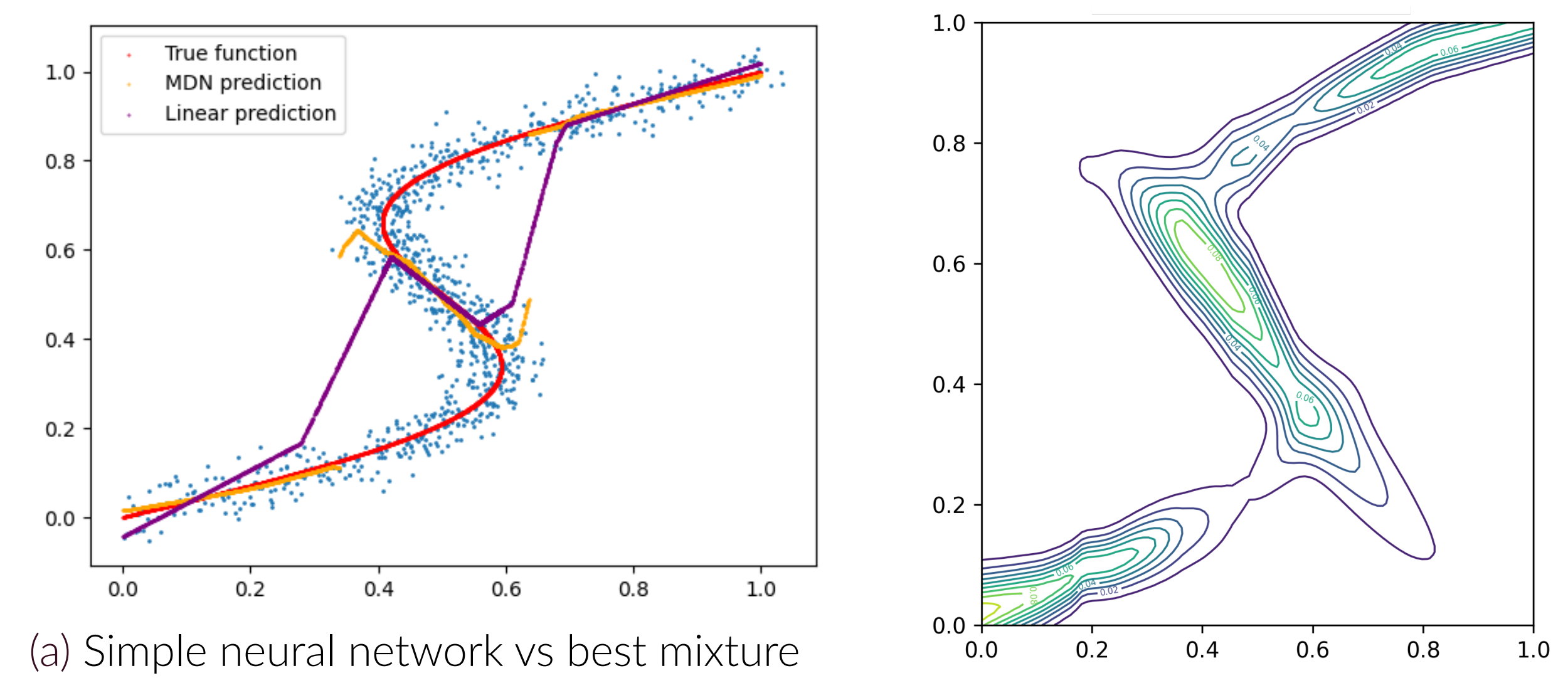
Why use Mixture Density Networks?

Mixture Density Networks offer a more nuanced approach by **capturing uncertainty and multi-modal data distributions** compared to models trained with simple MSE loss, which focuses solely on minimizing point-wise errors. It usage could be critical in works that need interpretability over the span of complex probability distributions of the data to leverage doubt on non-injectivty of processes.

Experiments : Toy dataset

Experiments with the inverse problem (non-injective) of

$$f(t) = t + 0.3\sin(2\pi t)$$



(a) Simple neural network vs best mixture for each point of a MDN with $m = 3$ on the hard toy dataset

(b) Contour plot of MDN prediction with $m = 3$ on the hard toy dataset

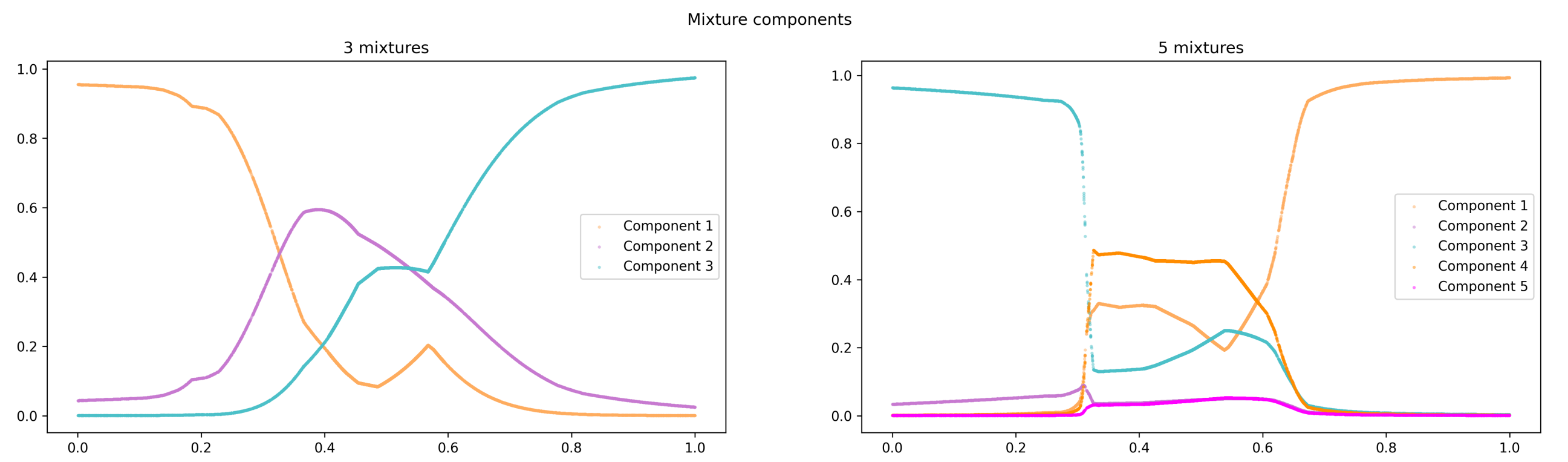


Figure 3. Values of α along the space for each mixture component of a model with $m = 3$ (left) and $m = 5$ (right) on the hard toy dataset.

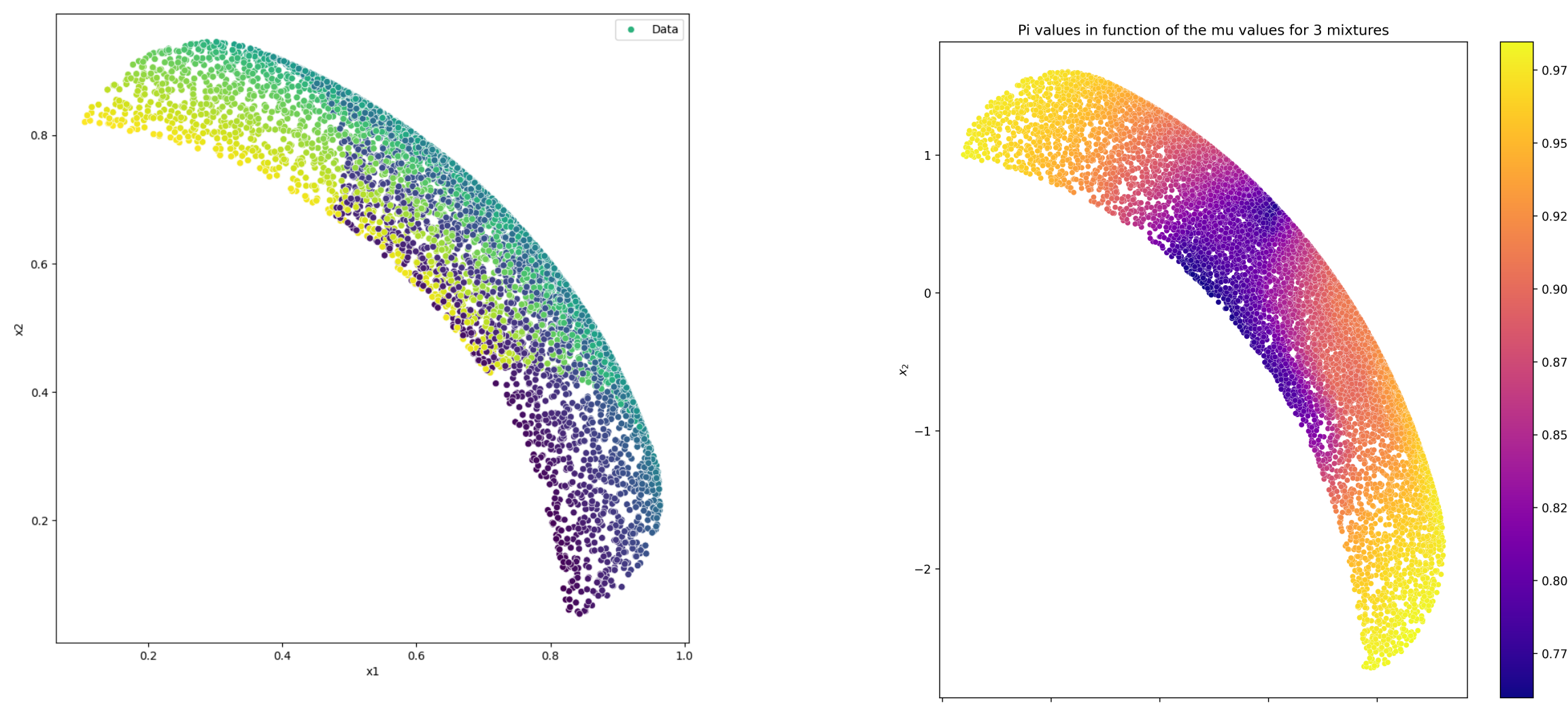
1. As anticipated, the linear neural network forecasts the average output of the ground truth, whereas the MDN accurately predicts a segment of the correct label.
2. The **optimal number of components** is directly linked to the number of modes in the output distribution.

Experiments : Robot kinematics

Experiments with the extremity (x_1, x_2) of a robot arm with two parts of length L_1, L_2 with angle θ_1 to the ground and θ_2 between the two parts.

$$x_1 = L_1\cos(\theta_1) - L_2\cos(\theta_1 + \theta_2)$$
$$x_2 = L_1\sin(\theta_1) - L_2\sin(\theta_1 + \theta_2)$$

where (θ_1, θ_2) in range $(0.3, 1.2) \times (\pi/2, 3\pi/2)$, $L_1 = 0.8$ and $L_2 = 0.2$.



(a) Overlaps of (x_1, x_2) generated by samples of (θ_1, θ_2) in range $(0.3, 1.2) \times (\pi/2, 3\pi/2)$

(b) Best α (also noted π) values for each points predicted

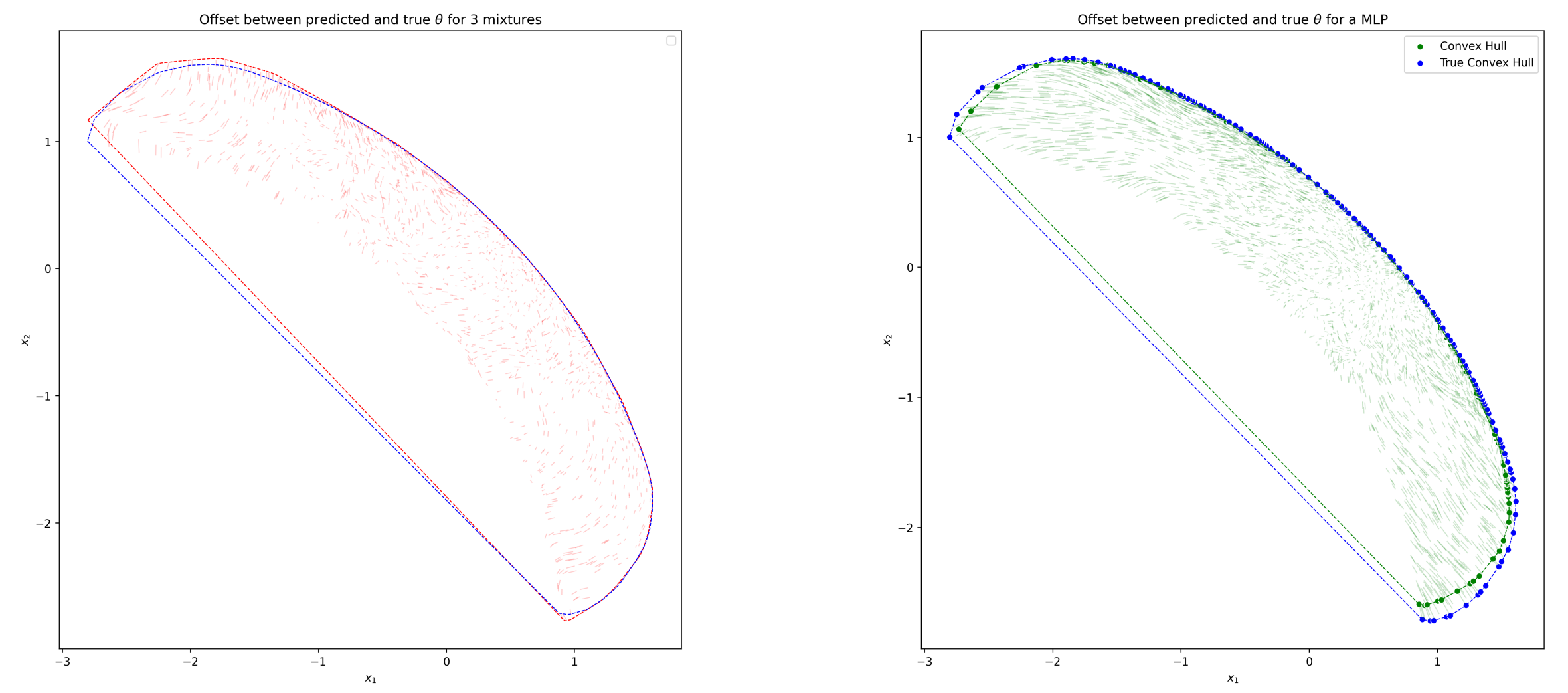


Figure 5. Positioning error as straight lines from desired positions to predicted ones with convex hulls

1. The MDN results in less positionning errors as it predicts the best mixture instead of the mean as for the linear model.
2. Important and interesting **implications in medical and robotic fields** where capturing incertainty is key.

Experiments : MNIST Regression

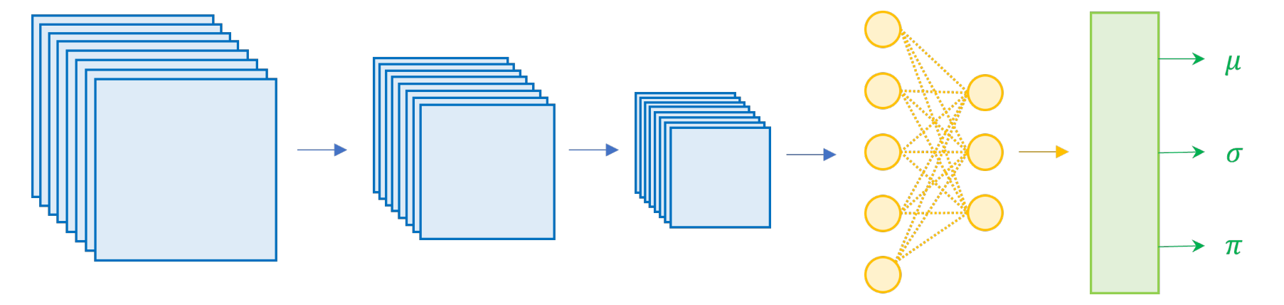
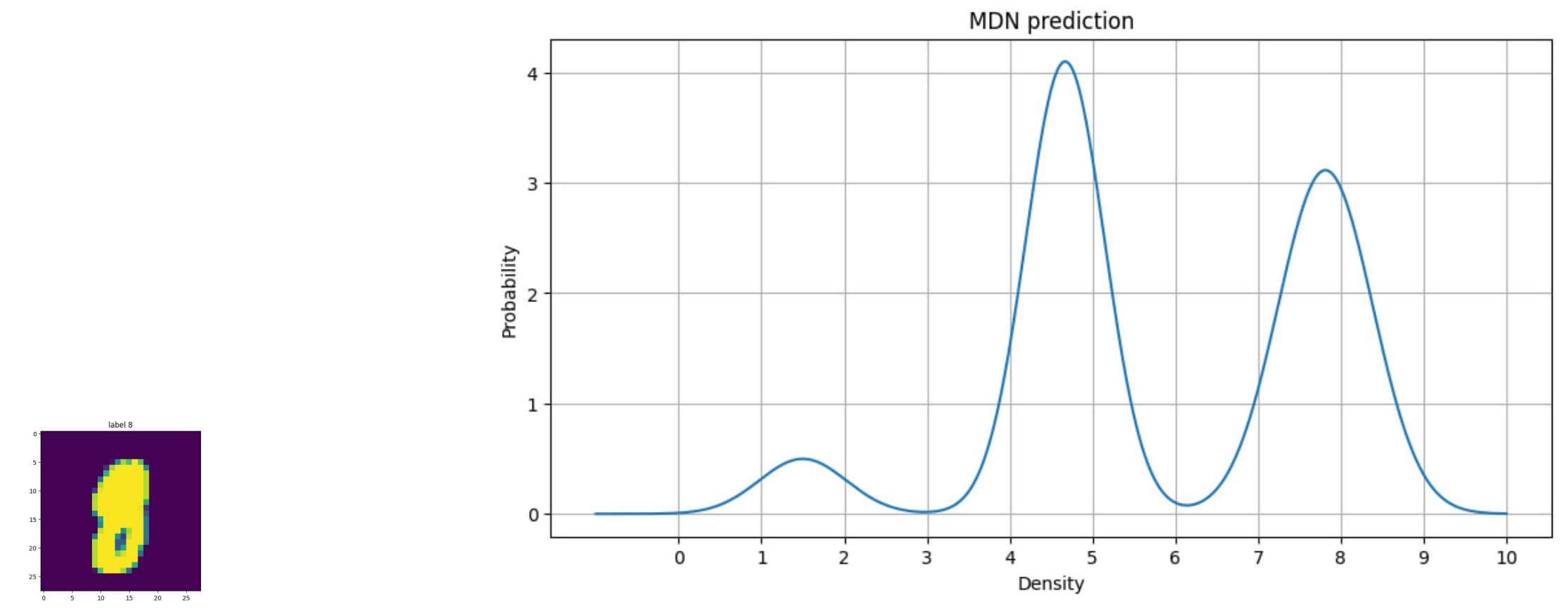


Figure 6. Convolutional Mixture Density Network composed of a standard convolutional network followed by a feed-forward network with the outputs sent into a Mixture Density head.



(a) MNIST sample

(b) MDN output on a sample of MNIST

The MDN surpasses the limitations of a simple neural network by offering multiple potential answers even when trained only with one correct label. We also demonstrate the compatibility of the MDN framework with more complex and deep neural network architectures.

References

[1] Christopher M. Bishop. Mixture density networks. Workingpaper, Aston University, 1994.

[2] Shlomo E. Chazan, Jacob Goldberger, and Sharon Gannot. Speech enhancement using a deep mixture of experts, 2017.

[3] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[5] Lara Orlandic, Tomas Teijeiro, and David Atienza. The coughvid crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. *Scientific Data*, 8(1):156, Jun 2021.

[6] Heiga Zen and Andrew Senior. Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3844–3848, 2014.

Poster session - Introduction to Probabilistic Graphical Models and Deep Generative Models

Master MVA 2023-2024