# Sparse Recurrent Mixture Density Networks for Forecasting High Variability Time Series with Confidence Estimates


TATA CONSULTANCY SERVICES

Narendhar Gugulothu, Easwar Subramanian and Sanjay P. Bhat

TCS Research, Hyderabad, India


## Introduction

- Proposed two variants of sparse Recurrent Mixture Density Networks (sparse RMDN) for forecasting, namely, **sparse LSTM-MDN** and **sparse ED-MDN**.
- Contains a feedforward layer which produces a sparse and low-dimensional representation of the high-dimensional input data.
  - The sparsity is achieved by imposing LASSO [1] penalty on the weights of the feedforward layer.
  - Each unit in the feedforward layer has access to only a subset of the input features [2], which results in unsupervised feature selection.
- The output of the feedforward layer is fed to the subsequent RNN (LSTM [3] or ED [4]) to capture temporal patterns.
- The output of the RNN is passed through Mixture Density Network (MDN) [5] to handle the variability in the data and to estimate the confidence of the forecast.

## Sparse LSTM-MDN

- $\mathbf{x} = (x_1, \cdots, x_t)$ denotes an input sequence of length $t$. Each $x_k \in \mathbb{R}^d, k \in \{1, \cdots, t\}$; $d$=input dimension.
- The model is required to provide a prediction $\mathbf{y}'_{t+1,\ldots,t+p}$ of $\mathbf{y}_{t+1,\ldots,t+p}$ given the input $\mathbf{x}_{1,\ldots,t}$.

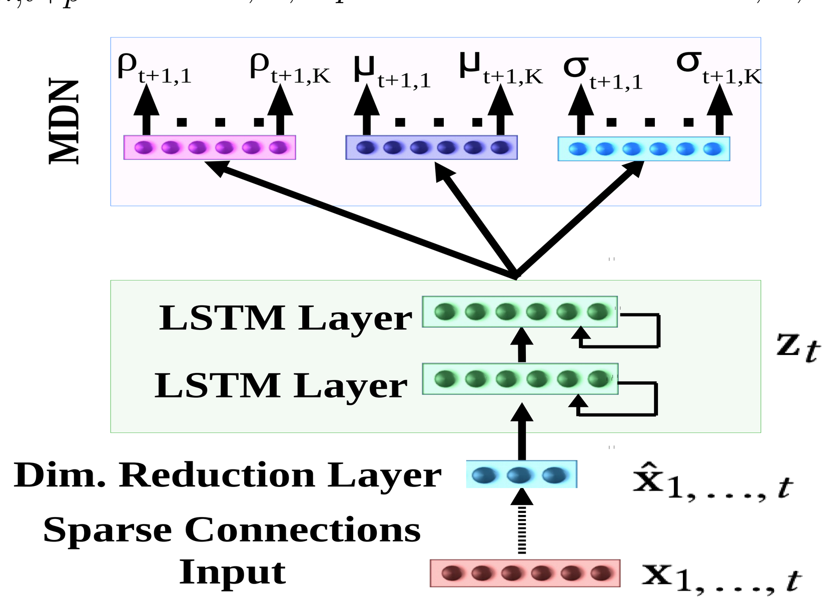


Figure: Proposed sparse LSTM-MDN for 1-step ahead forecasting with two LSTM layers and $K$ Gaussians.

- The output of the feedforward layer for input $\mathbf{x}_t^i$ is given by $\hat{\mathbf{x}}_t^i = f_{ReLU}(\mathbf{W}_f \cdot (\mathbf{x}_t^i)^T + \mathbf{b}_f)$.
  - $\mathbf{W}_f$ = Weights of the feedforward layer
- The intermediate term $\hat{\mathbf{x}}^i$ is then fed to subsequent LSTM layers.
- The output of the LSTM layers ($\mathbf{z}_t$) is fed to MDN to estimate the parameters of Gaussians.
- The parameters of $K$ Gaussian mixtures for a future time stamp $t'$ are estimated as follows:

$$\rho_{t',\cdot}(\mathbf{z}_t) = \mathrm{softmax}(\mathbf{W}_\rho \cdot \mathbf{z}_t + \mathbf{b}_\rho)$$
$$\mu_{t',\cdot}(\mathbf{z}_t) = \mathbf{W}_\mu \cdot \mathbf{z}_t + \mathbf{b}_\mu$$
$$\sigma_{t',\cdot}(\mathbf{z}_t) = \exp(\mathbf{W}_\sigma \cdot \mathbf{z}_t + \mathbf{b}_\sigma)$$

- The conditional distribution of predicting $\mathbf{y}_{t+1,\ldots,t+p}$ given $\mathbf{z}_t$ expressed as follows:

$$P(\mathbf{y}_{t+1,\ldots,t+p}|\mathbf{x}_{1,\ldots,t}; \mathbf{z}_t) = \prod_{t'=t+1}^{t+p} \sum_{k=1}^{K} \rho_{t',k}(\mathbf{z}_t) \mathcal{N}\Big(\mathbf{y}_{t'}; \mu_{t',k}(\mathbf{z}_t), \sigma_{t',k}(\mathbf{z}_t)\Big)$$

- The model parameters are learned by minimizing the negative log-likelihood of the distribution as shown below:

$$\mathcal{L}_{RMDN} = -\frac{1}{N} \sum_{i=1}^{N} \log P\Big(\mathbf{y}^i_{t+1,\ldots,t+p}|\mathbf{x}^i_{1,\ldots,t}; \mathbf{z}^i_t\Big)$$

  - $N$= Samples in train set
- The final loss function along with the Lasso penalty on $\mathbf{W}_f$ is thus given by

$$\mathcal{L} = \mathcal{L}_{RMDN} + \frac{\lambda}{d \times r} ||\mathbf{W}_f||_1$$

  - $r$ = Units in the feedforward layer and $r \leq \frac{d}{2}$
  - $\lambda$ = Regularization parameter, which controls sparsity level

## Sparse ED-MDN

- Sparse ED-MDN is a variant of sparse LSTM-MDN, where LSTM network is replaced by ED network.

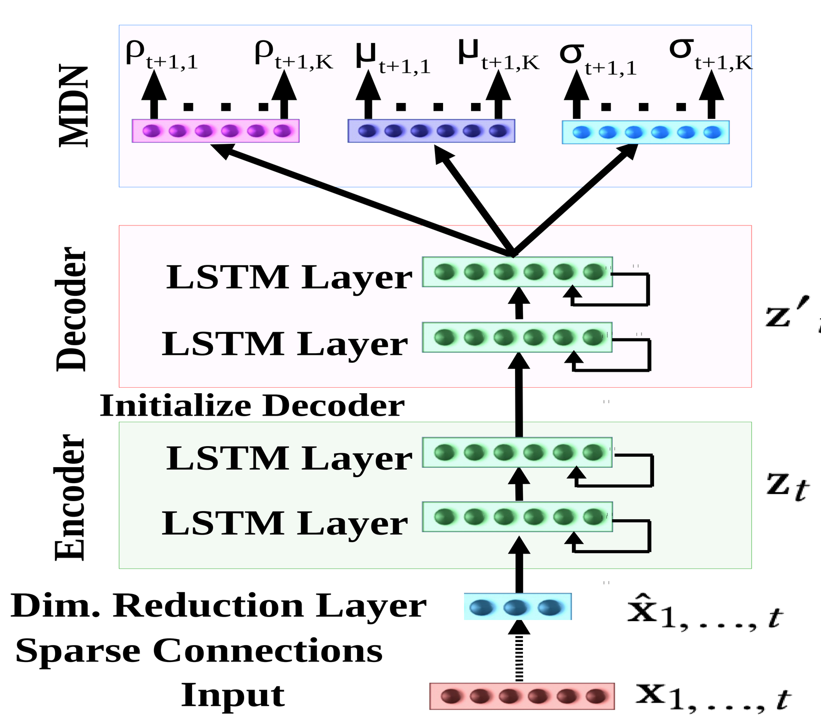


Figure: Proposed sparse ED-MDN.

## Results

Table: Performance comparison of proposed sparse RMDN based forecasting models on five datasets.

| | AEMO | | HomeA | | MedicalCenter-1 | | CentervilleHomes | | BrooksideHomes | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MSE | MAPE | MSE | MAPE | MSE | MAPE | MSE | MAPE | MSE | MAPE |
| **Standard LSTM** | 0.00159 | 9.03186 | 0.01182 | 46.98303 | 0.00559 | 17.71834 | 0.00159 | 12.6512 | 0.00276 | 30.47902 |
| **Standard ED** | 0.00237 | 11.03585 | 0.01172 | 44.10512 | 0.00586 | 17.5974 | 0.00165 | 13.35113 | 0.00275 | 29.24932 |
| **Sparse LSTM** | **0.00137** | **8.66642** | 0.01113 | 42.77446 | 0.00596 | 18.5316 | 0.00162 | 12.1824 | **0.00256** | 28.94780 |
| **Sparse ED** | 0.00170 | 9.25787 | 0.01177 | 43.95304 | 0.00619 | 18.5024 | 0.00154 | 13.02479 | 0.00273 | 31.45647 |
| **LSTM-MDN** | 0.00227 | 10.36923 | 0.01295 | 28.36924 | 0.00559 | **17.22558** | 0.00157 | 12.06838 | 0.0028 | 28.93074 |
| **ED-MDN** | 0.00199 | 9.37978 | 0.01381 | 35.16026 | 0.00587 | 17.26241 | 0.00155 | 12.17745 | 0.00277 | 28.51284 |
| **Sparse LSTM-MDN** | 0.00167 | 9.14346 | 0.01188 | **25.64572** | 0.00553 | 18.54566 | **0.00150** | **11.96106** | 0.00281 | **26.81967** |
| **Sparse ED-MDN** | 0.00176 | 9.19194 | **0.01170** | 29.37821 | **0.00536** | 18.84936 | 0.00153 | 12.42076 | 0.00299 | 27.03677 |
| **Ensemble** | **0.00134** | **8.07125** | **0.01015** | 34.10546 | **0.00510** | **17.11900** | **0.00139** | **11.74229** | **0.00242** | 27.27441 |

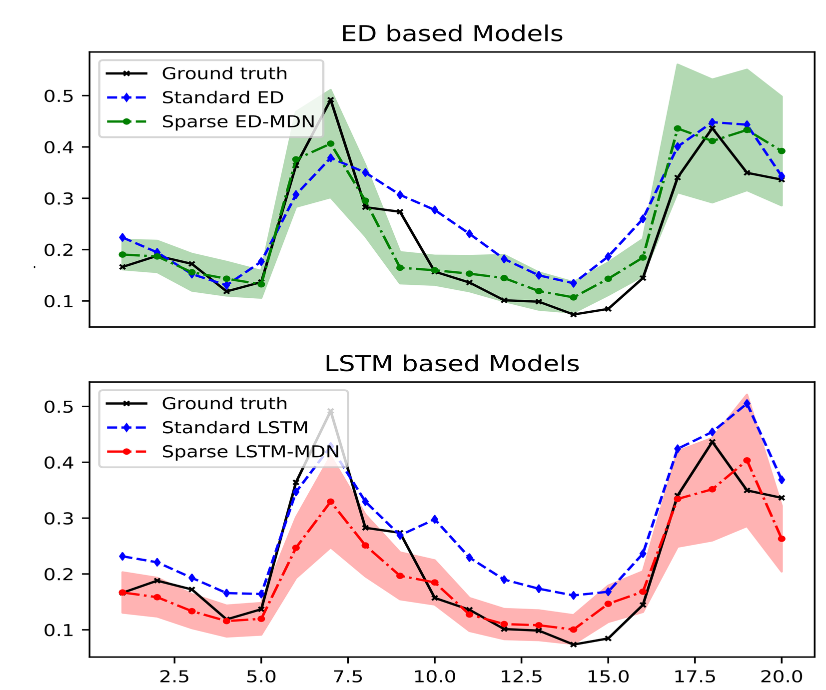


Figure: HomeA

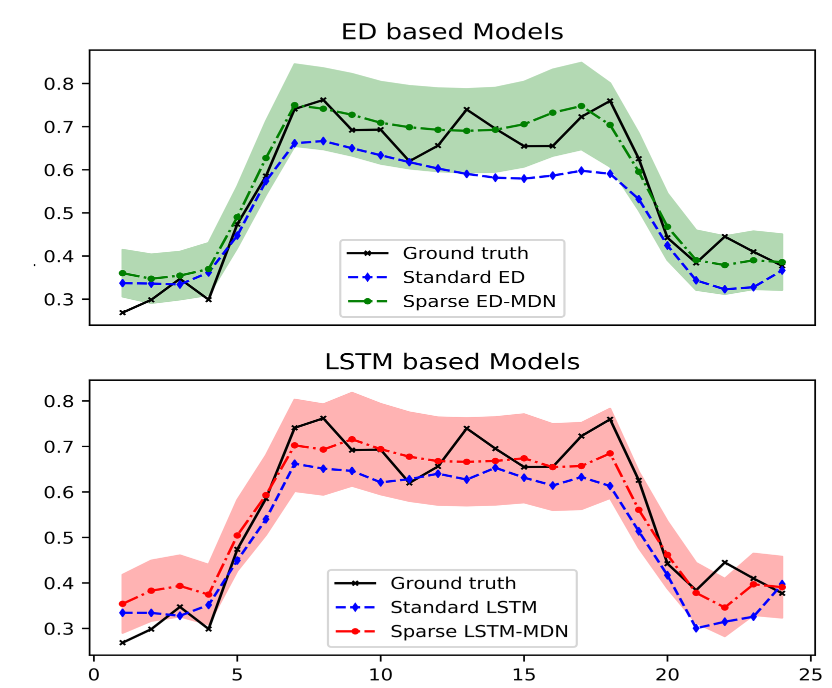


Figure: MedicalCenter-1

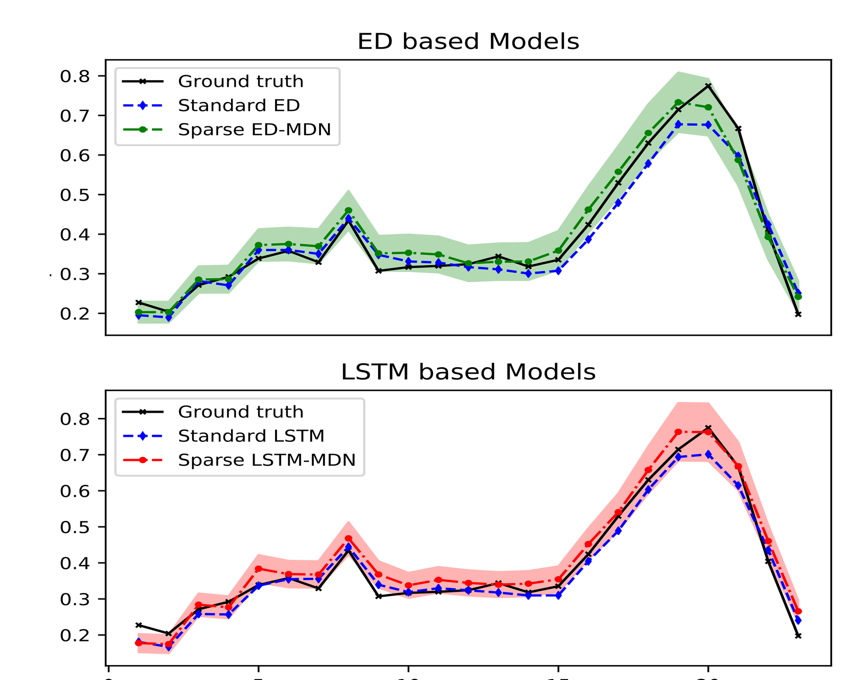


Figure: CentervilleHomes

Figure: Comparison of proposed sparse RMDNs and standard RNN based models with the ground truth for forecasting[a].

## Observations

- Sparse LSTM and sparse ED outperformed their non-sparse counter parts.
- MDN variants of LSTM and ED are better than standard LSTM and ED.
- Sparse RMDN based forecasting models outperformed all other approaches.
- Sparse RMDNs are superior in capturing the variability and handling high-dimensional data than their non-sparse, non-MDN counterparts.
- Estimated confidence $\sigma$ is low whenever the prediction error is low and it is high otherwise.

[a] Refer paper for plots on AEMO and BrooksideHomes datasets.

## Conclusion

- Proposed sparse RMDN models outperforms existing RNN based models on forecasting.
- The sparse RMDN models perform point-wise dimensionality reduction and unsupervised feature selection of high-dimensional data.
- Captures temporal patterns using underlying RNN architectures.
- Handles high variability and trend shifts better using MDN.
- Provides a confidence estimate of the forecast.

## References


[1] Robert Tibshirani.
Regression shrinkage and selection via the lasso.
*Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[2] Narendhar Gugulothu, Pankaj Malhotra, Lovekesh Vig, and Gautam Shroff.
Sparse neural networks for anomaly detection in high-dimensional time series.
In *AI4IOT workshop, IJCAI*, 07 2018.

[3] S. Hochreiter and J. Schmidhuber.
Long short-term memory.
*Neural computation*, 9(8):1735–1780, 1997.

[4] Ilya Sutskever, Oriol Vinyals, and Quoc V Le.
Sequence to sequence learning with neural networks.
In *NIPS*, pages 3104–3112, 2014.

[5] Christopher M Bishop.
Mixture Density Networks.
Technical report, Citeseer, 1994.


## Contact Information


Email: narendhar.g@tcs.com
WhatsApp: +91-9705902716