SEMI-SUPERVISED TIME-SERIES LEARNING

A PREPRINT

Nikita Klyuchnikov
Skoltech
nikita.klyuchnikov@skoltech.ru

Rodrigo Rivera
Skoltech
rodrigo.riveracastro@skoltech.ru

October 27, 2018

ABSTRACT

This work is dedicated to semi-supervised approaches for learning time series. We have conducted an exploratory analysis with clustering methods of two real datasets and applied a (semi-supervised) Variational Auto-Encoders on them. The latter has also been studied in more detail on a toy artificial dataset with two-variate sine-waves. The results show that the surveyed method is not adequate for time series data, specially for clustering tasks.

***K*eywords** Semi-supervised, VAE, time series, learning

Contribution of Authors

- Nikita Klyuchnikov: Implementation and experiments with semi-supervised VAE, report preparation
- Rodrigo Rivera: Exploratory data analysis with clustering methods, report preparation

1 Introduction

Learning representations from data is the quintessential task in machine learning. Multiple models of different characteristics have been proposed over many decades, a recent one, based on the concept of Variational Autoencoders, Kingma and Welling (2014), has found significant success across a myriad of contexts and tasks. In this report, we evaluate a further development of this model focusing on semi-supervised learning.

2 Methods

For this analysis, it was decided to make use of a series of methods for analysis and learning. A brief overview is presented below. The reader is advised to consult the reference material for a detailed presentation of the material.

Variational Auto-Encoder
Ordinary Variational Auto-Encoder (VAE) was proposed by Kingma and Welling (2014). The idea of this auto-encoding model is to approximate posterior distribution of latent parameters z with Gaussian inference network $q(z|x)$ and use it for evidence lower bound optimization. Whereas the prior is also assumed to be Gaussian $p(z) \sim \mathcal{N}(0, I)$.

Semi-Supervised Variational Auto-encoder
Semi-supervised Learning with Deep Generative Models was proposed by Kingma et al. (2014), who made the following modifications to ordinary VAE:

- the dataset consists of two parts: items with supervised labels $D = \{(x_i, y^i)\}_{i=1}^{n_s}$ and unlabeled items $X = \{x_i^{(u)}\}_{i=1}^{n_u}$, for which labels are considered as additional latent variables;
- labels are generated from some categorical distribution $\text{Cat}(y|\pi)$;
- approximate posterior factorizes on Gaussian and multinational distributions of z and y:

$$q(z, y|x) = \mathcal{N}\left(z|\mu(x), \text{diag}(\sigma^2(x))\right) q(y|\pi(x))$$

These modifications lead to different training scheme, in particular, to new evidence lower bound:

$$\mathcal{J}_\alpha = \sum_{x \in X} \mathcal{U}(x) + \sum_{(x,y) \in D} \left[\mathcal{L}(x, y) - \alpha \log q(z, y|x)\right], \tag{1}$$

where $\mathcal{U}(x) = \sum_y \left[q(y|x)\left(L(x, y) - \mathbb{H}(q(y|x))\right)\right]$ and

$$L(x, y) = \mathbb{H}(q(z|x, y)) - \mathbb{E}_{q(z|x,y)}\left[\log p(x|y, z) + \log p(y) + \log p(z)\right].$$

Original paper studied the model performance on images, in this work we adjusted the model for time-series: the key difference will be in the amortized inference networks architecture, that represent $\mu(x)$, $\sigma^2(x)$ and $\pi(x)$ (see the particular architecutre that we implemented in Figure 21).

K-shape
Proposed by Paparrizos and Gravano (2015). K-Shape relies on an iterative refinement procedure, which creates homogeneous clusters using a normalized version of the cross-correlation measure as its distance measure in order to consider the similarity of two time series based on their shapes to compare them. It also computes cluster centroids, which are used in every iteration to update the assignment of time series to clusters.

Soft-DTW
A work of Cuturi and Blondel (2017) proposing a smooth version of DTW. A method that finds the optimal non-linear alignment between two time series. The traditional DTW considers two time series $Q = \{q_1, \ldots, q_n\}$ and $C = \{c_1, \ldots, c_n\}$ of the same length n. A $n \times n$ matrix is constructed, whose i, j^{th} element are the Euclidean distance between the elements q_i and c_j. The objective is to find an optimal aligment between both time series by a path through the matrix W, which minimizes the cumulative distance path. Thus, it minimizes the Euclidean distance $W^* = argmin_W(\sqrt{\Sigma_{k=1}^K w_k})$, where w_k is an element of the matrix W containing the distances between the points.

K-Means with Fast Global Alignment Kernel Presented originally by Cuturi (2011), it proposes a reformulation of DTW as toeplitz kernel. The authors show results that not only offer significantly improvements in reduction of the test error but also a faster computation.

Table 1: Overview of data sets used for experiments

Data set	Elements	Periods	Characteristics
Inventory demand	2400	44	Representing demand for 800+ products across 4 delivery periods. Heterogeneous demand.
MWD dataset	31352	720	Represents physical characteristics of oil-wells drilling at various stages. Long 4-d time series logs from 60 wells. Slices with length 180 periods each

Gaussian Mixture Model
Gaussian Mixture Model (GMM) is a commonly used method for clustering and density estimation. It assumes that all generated data points are derived from a mixture of a finite Gaussian distribution with unknown parameters. The parameters are often derived with methods such as maximum a posteriori (MAP).

LB-Keogh
It is a popular method proposed by Keogh and Ratanamahatana (2005), to speed dynamic time warping (DTW), as the LB Keogh lower bound is linear whereas the commonly used DTW has a quadratic complexity. More concretely, LB-Keogh is defined in equation 2

$$LBKeogh(Q,C) = \Sigma_{i=1}^{n}(c_i - U_i)^2 I(c_i > U_i) + (c_i - L_i)^2 I(i_i < L_i) \tag{2}$$

where U_i and L_i are upper and lower bounds for time series Q which are defined as $U_i = max(q_{i-r} : q_{i+r})$ and $L_i = min(q_{i-r} : q_{i+r})$ for a reach r and $I(\cdot)$ is the indicator function.

Connection between methods and surveyed model
The main driver between the choice of methods during the exploratory phase and the surveyed model is the need to identify efficient methods to learn representations of time series. Due to their simplicity, time series often tend to lack context and additional information (i.e., features). This poses a challenge when one has access to a group of time series belonging to the same generating process but displaying diverse behaviors. The ability to identify similar time series or methods that yield relevant information that improves the task of time series prediction is thus very valuable. The surveyed method, due to its semi-supervised nature, can be considered for clustering. Thus, the clustering methods evaluated here serve not only to assess the suitability of the data sets used but also as a comparison.

3 Exploratory data analysis

Before carrying out any learning task, a set of data sets were chosen to evaluate the suitability of the observed method for time series. On one side, a toy data set that we call *'MNIST Time series'* was generated consisting of a positive, see figure 15b, and a negative, see figure 15a, class. On the other side, two real data sets from commercial partners were selected. Among other reasons, a motivating factor was that they are diametrically different. One of them has long time series, consisting of 720 periods, whereas the other one consists of 44. Further, one is multi-dimensional and the other uni-dimensional. An overview of both time series is shown below in table 1.

One of the first tasks to assess if the data sets are viable to be used for a learning task was to use them for data clustering. A motivation is that the semi-supervised approach can generate predictions based on labels; a process akin to a 'hard clustering'. In addition, the heterogeneous characteristics of the data sets and the existence of corner cases (i.e., stable time series with sudden spikes) were reasons to believe that the data would pose a challenge to the semi-supervised VAE method evaluated in this work. Thus, a group of clustering algorithms were selected among the state of the art. They are summarized in table 2 and presented in section 2.

$$AdjustedRandScore = \frac{(RandScore - ExpectedRandIndex)}{[max(RandIndex) - ExpectedRandIndex]} \tag{3}$$

Further, to evaluate the different clustering methods, it was decided to use the Adjusted Rand Score as an evaluation metric, see equation 3. This is a popular metric to compare the results of a clustering task. It assess the similarities between two data clusters using the Rand Index. For a further description, the reader is invited to consult Rand (1971).

Table 2: Overview of clustering methods used

Cluster method	Data set	Nr. Clusters	Notes
K-shape	Inventory demand	20	Gave the best results from all methods
K-means Kernel	Inventory demand	20	Global alignment kernel with 10 iterations. Very slow.
Soft DTW	Inventory demand	20	Method differentiable everywhere, using soft minimum gradient computed in quadratic time
GMM	Inventory demand	5	It used LB-Keogh as a distance. The results are then used to fit the GMM
K-shape	MWD	5	Slightly better results than the other methods, but with the advantage of faster computational times
Soft DTW	MWD	5	Faster than regular DTW and with similar results, but inferion vis a vis K-shape
GMM	MWD	5	

Methodology
For each method proposed, a series of clusters were computed. Out of this, a centroid was obtained. This centroid is another time series. Together with the time series belonging to each cluster, the Adjusted Rand Index was calculated. This can be seen for example in figure 10 with the results of this distance metric, a boxplot was generated to visualize the results, for example 2. The results of this exercise are discussed in section 5

4 Results

Toy sine-waves
In this section we considered an artificial dataset with time-series in the form of sine-waves with different phases and periods. The second channel is defined by the class label: for positive class the second channel is identical to the first one (see example in Figure 15b), for negative class the second channel is inverted (negation) of the first one (see example in Figure 15a).

We studied 5 various ways to fit sin-waves data: completely-/semi-/weakly- supervised modes actually refer to our semi-supervised VAE model with 2-dimensional latent space for time-series, however, these modes reflect the share of labels provided to the model 100%, 50% and 0% respectively. Other two methods are ordinary VAE with the same latent dimensionality and the extended one by the number of labels (i.e. 4-dimensional). We run 5 times each model with different seeds and plotted mean squared error (MSE) between representations and original time-series in Figure 20, learned latent spaces by semi-supervised and unsupervised models are shown in figures 18a and 18b. Several examples with reconstructions are shown in Figure 19. Latent spaces with fixed positive and negative labels are represented in figures 16 and 17 respectively.

Real datasets
Examples of time-series from MWD and Inventory demand datasets with corresponding reconstructions by semi-supervised VAE are provided in figure 23. Training progress is represented by MSE (on a hold-out sample) in figure 22. With these data sets, it was confirmed the unsuitability of the surveyed method for learning tasks with time series. Section 5 discussed this with further detail.

5 Discussion

The objective of this project was to assess if a semi-supervised approach is well-suited to learn representations of time series, as part of the validation process it was decided to carry out clustering tasks as described on the previous sections. The results were however discouraging both for clustering and semi-supervised learning.

Clustering
The clusters had very low Adjusted Rand Scores, thus indicating that the time series had little similarities among them. The results of the clustering task were therefore disappointing. This can be explained in part due to the high number

of time series used and their length. This was exacerbated by the low number of clusters selected. With few clusters available and many time series to cluster, it is to be expected that the clusters will be of heterogeneous nature. At the same time, for most methods it is not realistic to compute a large number of clusters. During the experiments, it was noticeable that most clustering methods, even those among the state of the art, are not well-suited for lengthy time series or for large amount of them. Although all methods gave poor results, as it can be seen on the box plots included in the Appendix, K-shape had the positive aspect of being fast. Therefore, one takeaway from the experiment is to consider K-shape as the go-to candidate for benchmarking clustering tasks.

Semi-Supervised Deep Generative Model
Although initially promising, further inspection and study of the model revealed that it is not well-suited for learning tasks with time series. On one side, its nature requires the presence of labels, this is already a limitation for time series data. On the other, the method is not adequate when a semi-large (i.e., more than 50) number of labels are available. In addition, the model requires variance within the time series. For situations where the time series consists of mostly constant values, the method is unable to learn. This can be subdued by perturbing the time series with Gaussian noise and transforming it with a log-transform or similar. However, this is not a silver bullet either nor it compensates significantly. It can be speculated that for medium-sized data sets of time series with moderate length and variance, the method might give reasonable results. Yet, with the selected data sets, it was not possible to confirm this.

6 Conclusion

This projected consisted on exploring if learning representations of time series can be accomplished using variational auto-encoders (VAE). Although at first promising, the results were not satisfactory. In fact, this research argues that VAE is suboptimal for this type of data structure, specially for the case of clustering of time series. Here, it became evidently, that any run-off-the-mill clustering model achieves better results than the semi-supervised VAE model discussed in this work. This research proposes instead to choose K-Shape as the go-to method for time series clustering. For general learning tasks, the surveyed method also did not perform adequately, specially if the time series consisted of constant values. In conclusion, although VAE has shown strong results in various settings and the semi-supervised flavor makes the model even more appealing, for time series related tasks, this is not the righ model of choice.

References

Cuturi, M. (2011). Fast global alignment kernels. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 929–936.

Cuturi, M. and Blondel, M. (2017). Soft-dtw: a differentiable loss function for time-series. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 894–903.

Keogh, E. and Ratanamahatana, C. A. (2005). Exact indexing of dynamic time warping. *Knowl. Inf. Syst.*, 7(3):358–386.

Kingma, D. P., Mohamed, S., Rezende, D. J., and Welling, M. (2014). Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pages 3581–3589.

Kingma, D. P. and Welling, M. (2014). Stochastic gradient vb and the variational auto-encoder. In *Second International Conference on Learning Representations, ICLR*.

Paparrizos, J. and Gravano, L. (2015). k-shape: Efficient and accurate clustering of time series. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, SIGMOD '15, pages 1855–1870, New York, NY, USA. ACM.

Rand, W. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850.

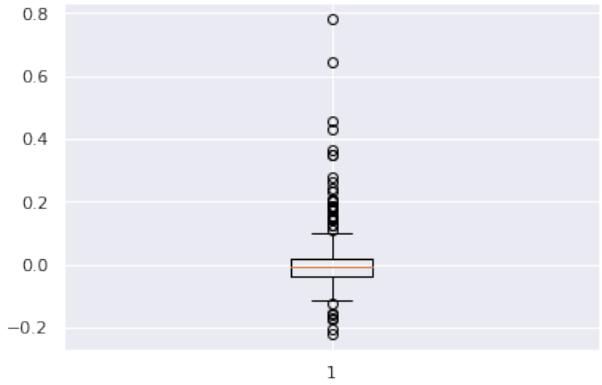


Figure 1: Boxplot of adjusted rand score of Gaussian Mixture Model for inventory demand

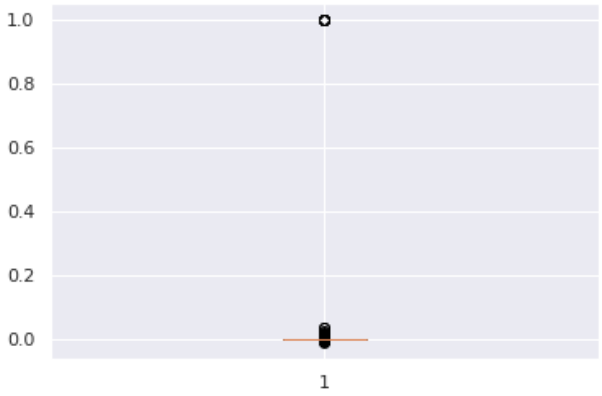


Figure 2: Boxplot of adjusted rand score of K-Shape clustering for inventory demand data set

A Appendix

Following figures were used accross this work

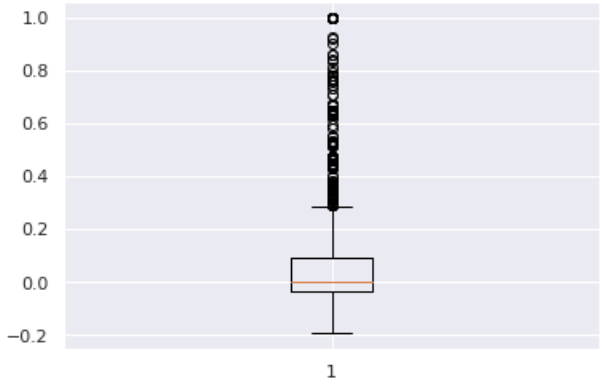


Figure 3: Boxplot of adjusted rand score of K-Means Kernel clustering for inventory demand data set

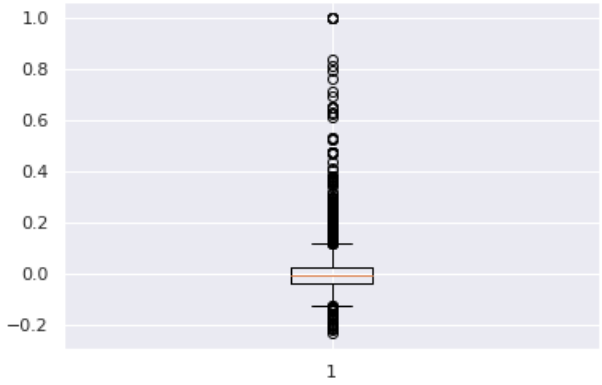


Figure 4: Boxplot of adjusted rand score of Soft-DTW Kernel clustering for inventory demand data set

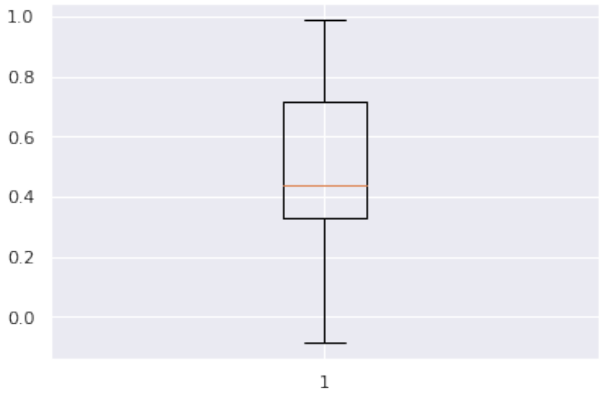


Figure 5: Boxplot of adjusted rand score of Gaussian Mixture Model for MWD data set

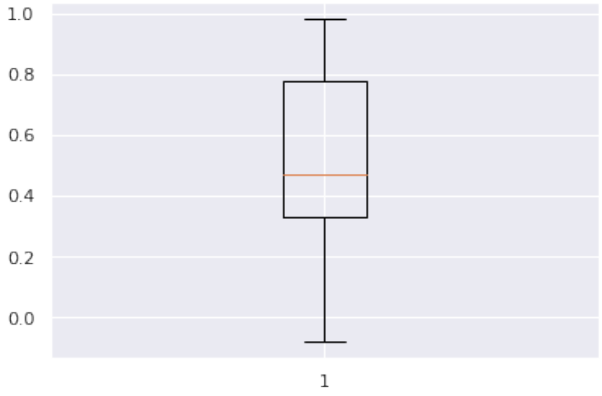


Figure 6: Boxplot of adjusted rand score of K-Shape clustering for MWD data set

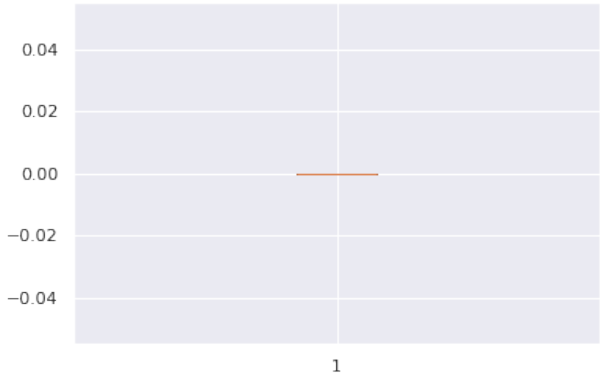


Figure 7: Boxplot of adjusted rand score of K-Means Kernel clustering for MWD data set

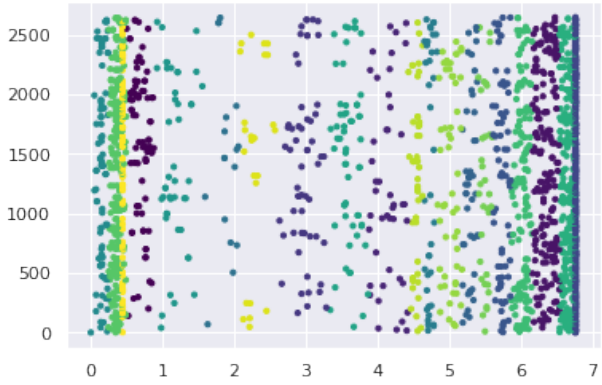


Figure 8: Clusters using Gaussian Mixture Model for inventory demand

A.1 Box Plots

A.2 Clusters

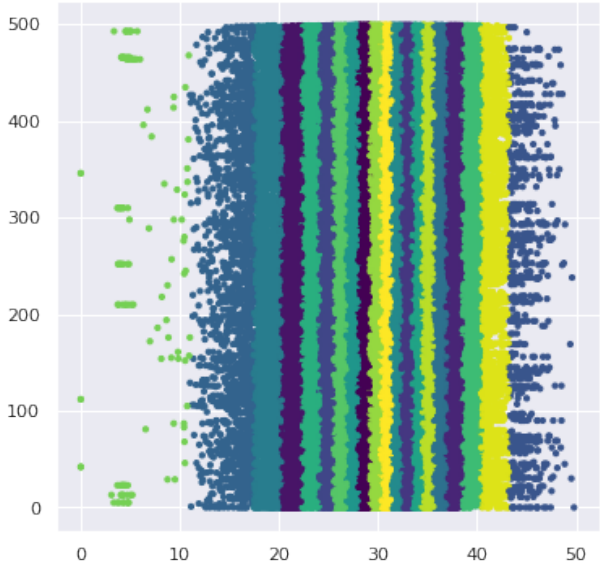


Figure 9: Clusters using Gaussian Mixture Model for MWD data set

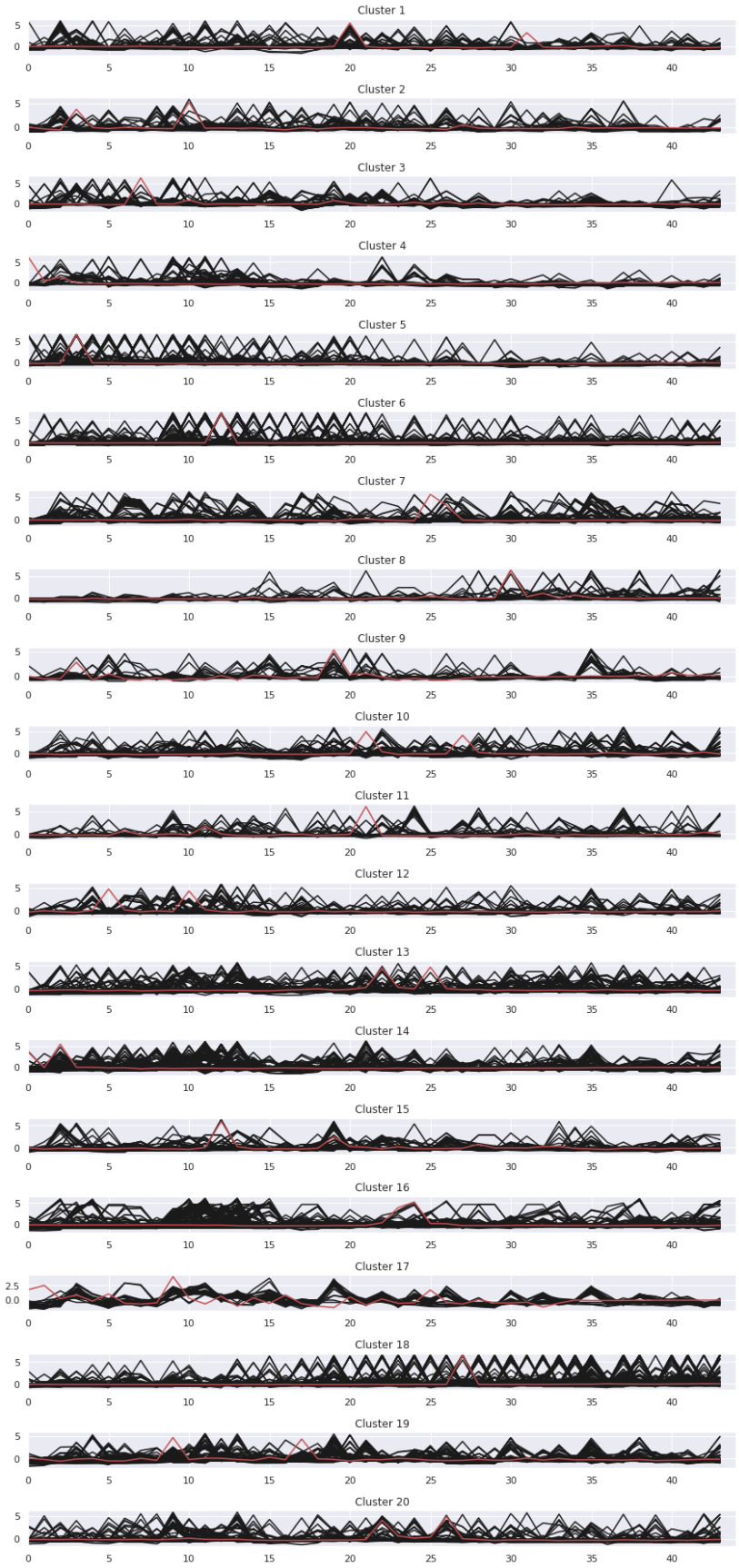


Figure 10: Clusters using K-Shape clustering for inventory demand data set

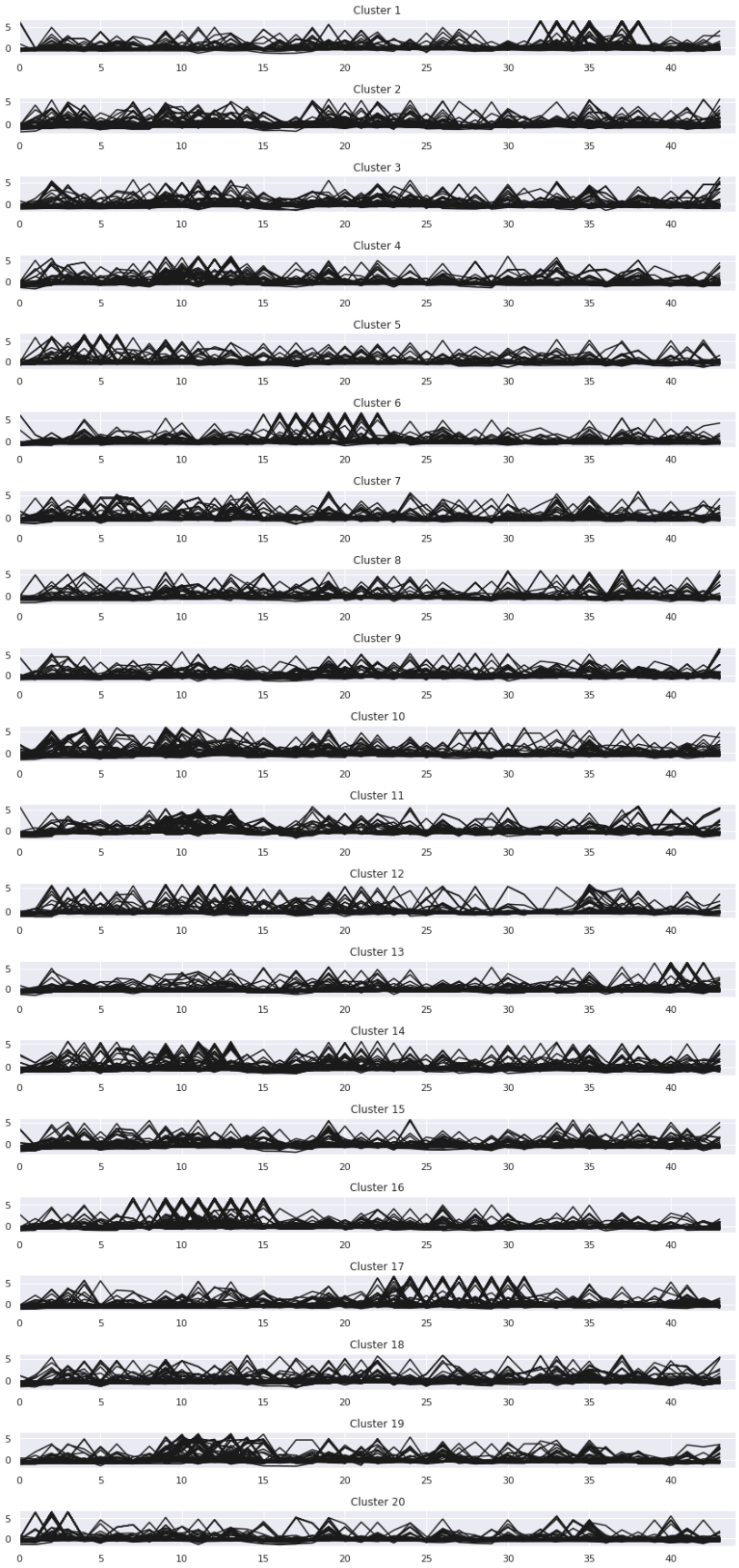


Figure 11: Clusters using K-Means Kernel clustering for inventory demand data set

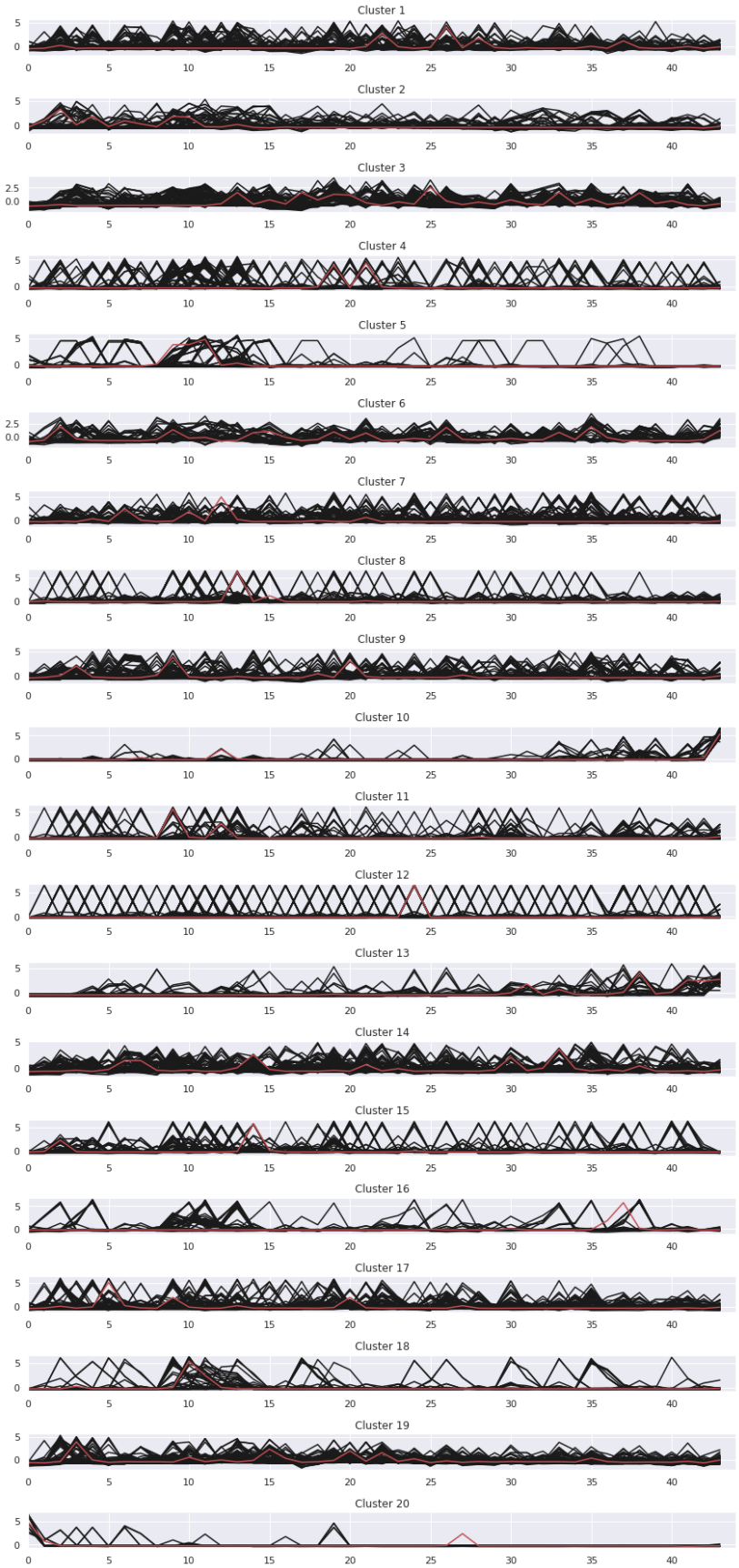


Figure 12: Boxplot of adjusted rand score of Soft-DTW Kernel clustering for inventory demand data set

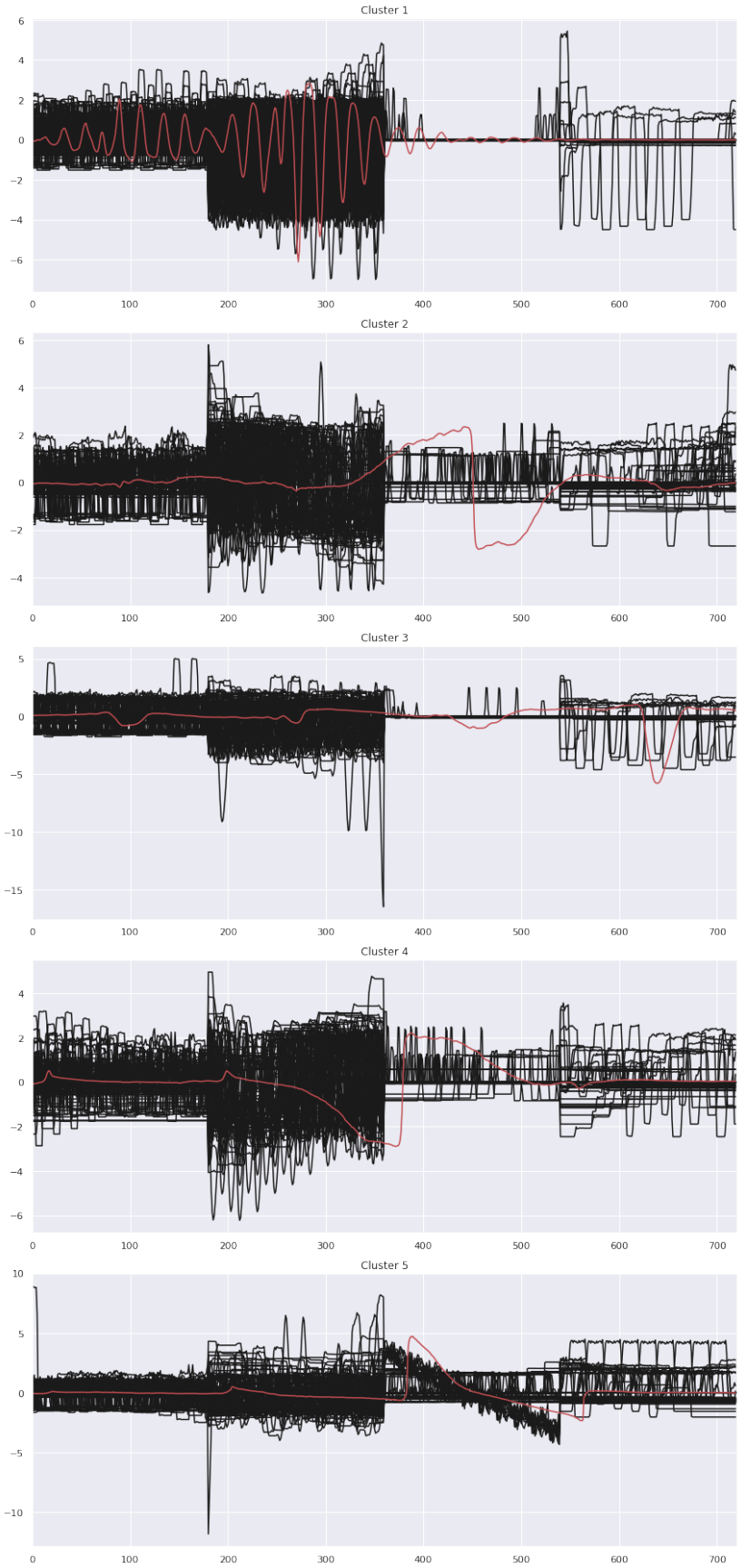


Figure 13: Clusters using K-Shape clustering for MWD data set

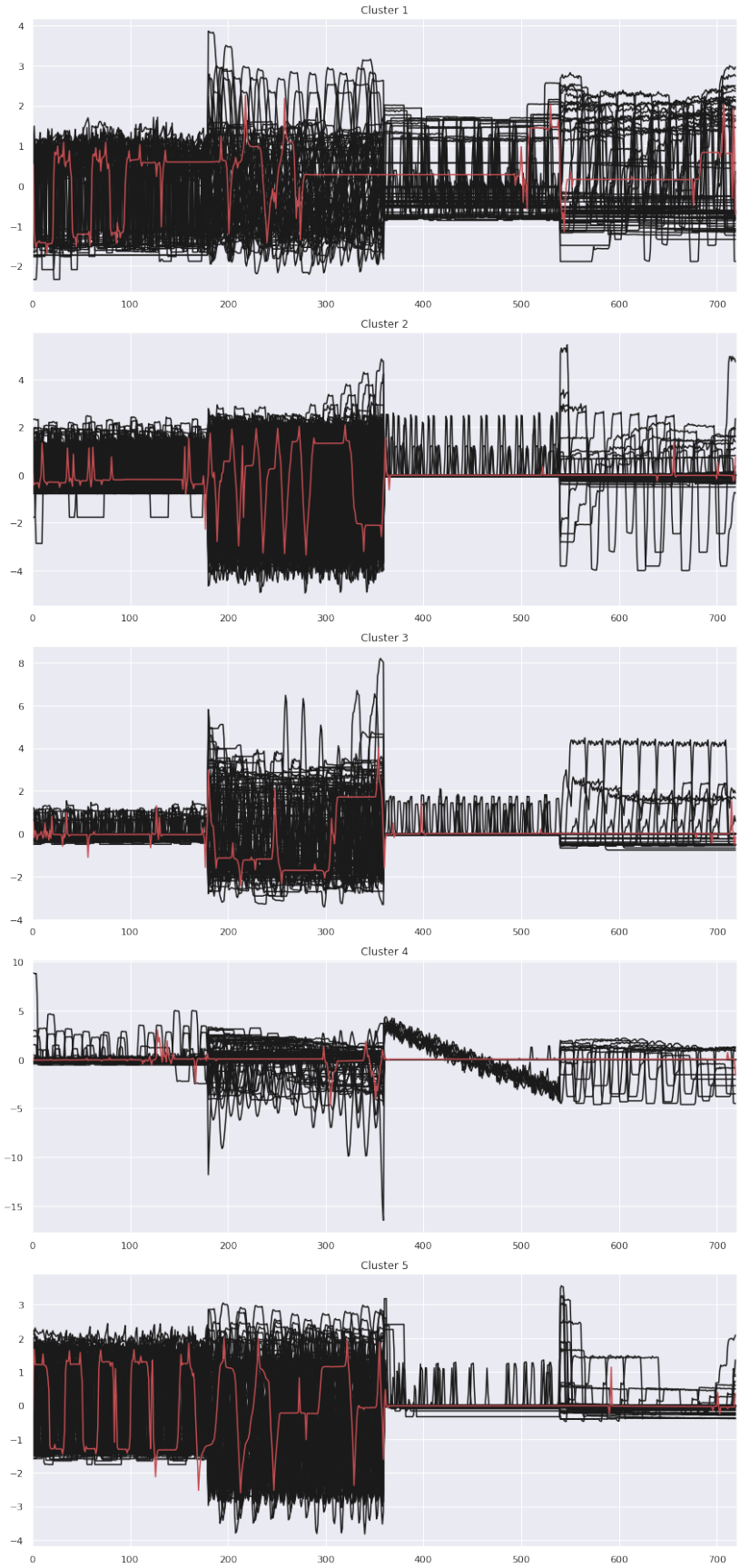
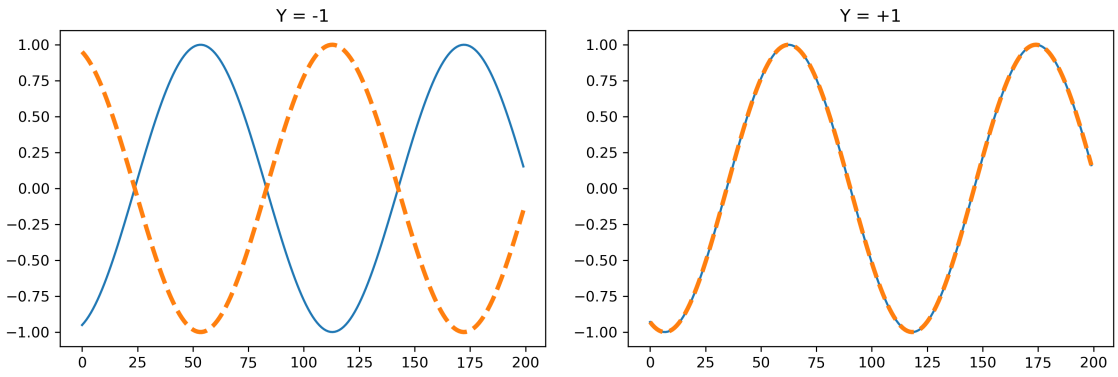


Figure 14: Clusters using K-Means Kernel clustering for MWD data set



(a) Negative class example.

(b) Positive class example.

Figure 15: Examples of sine-waves. Blue curve corresponds to the first channel, orange dashed curve corresponds to the second channel.

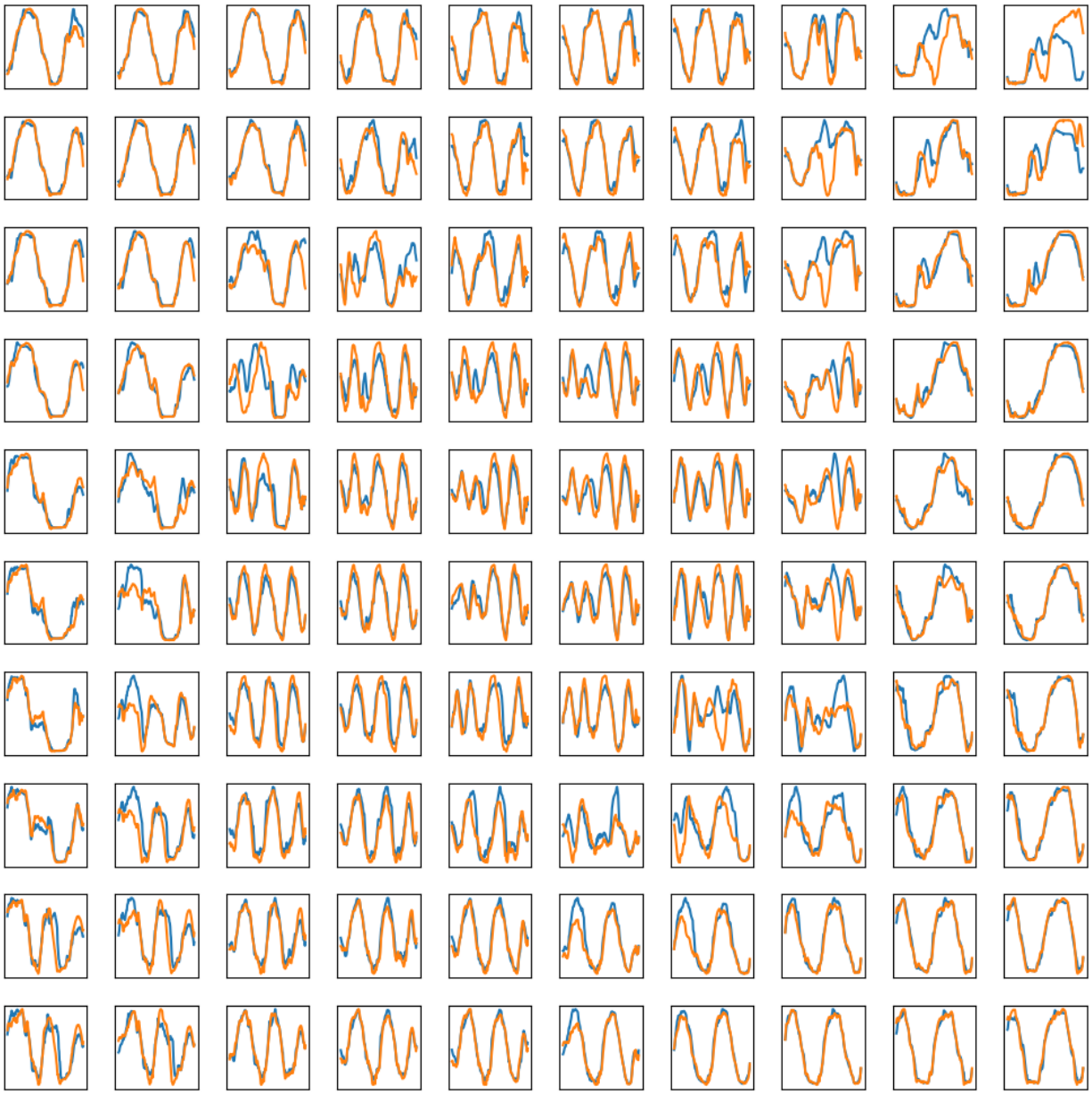


Figure 16: Exploration of latent space with positive label.

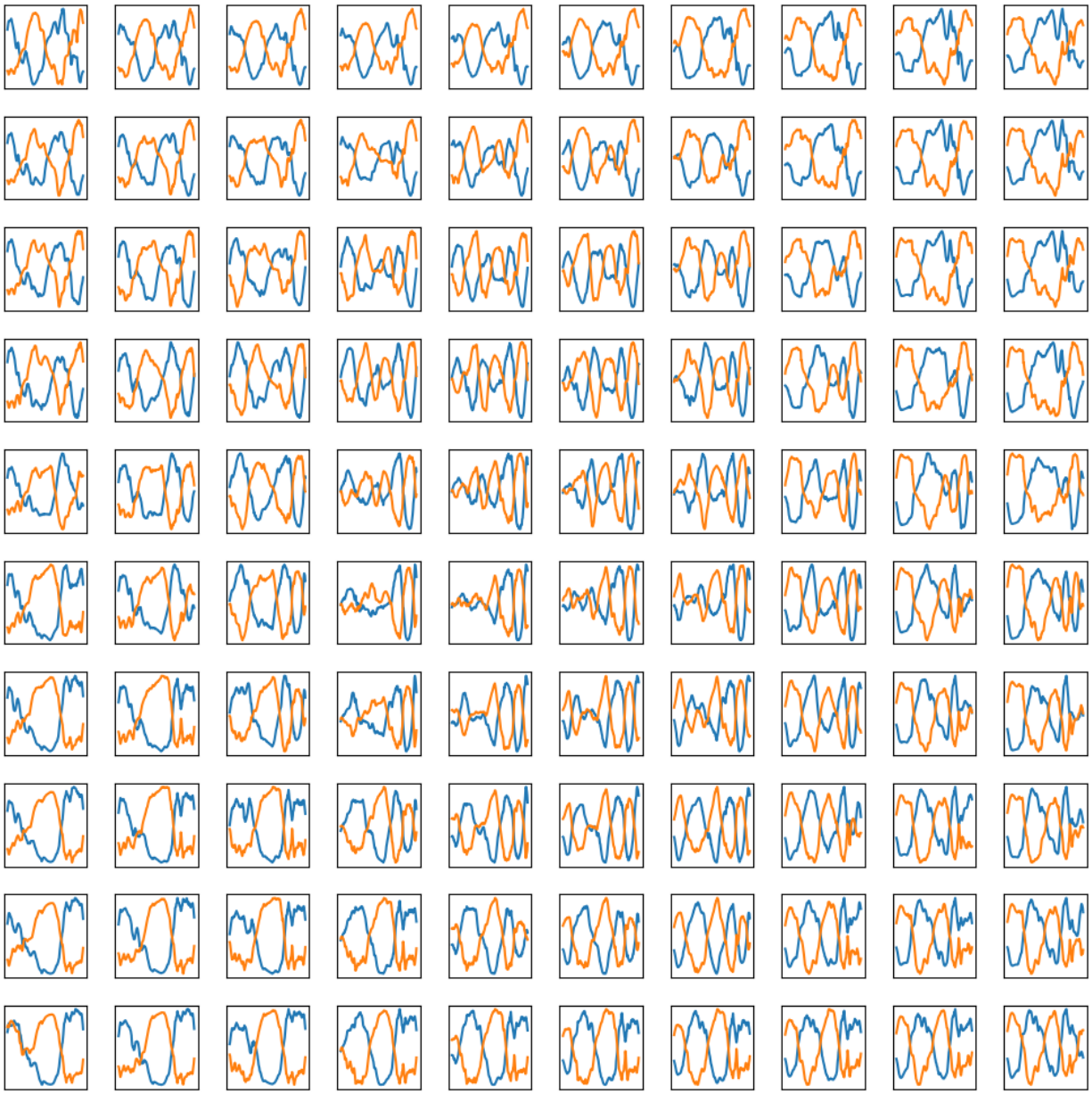
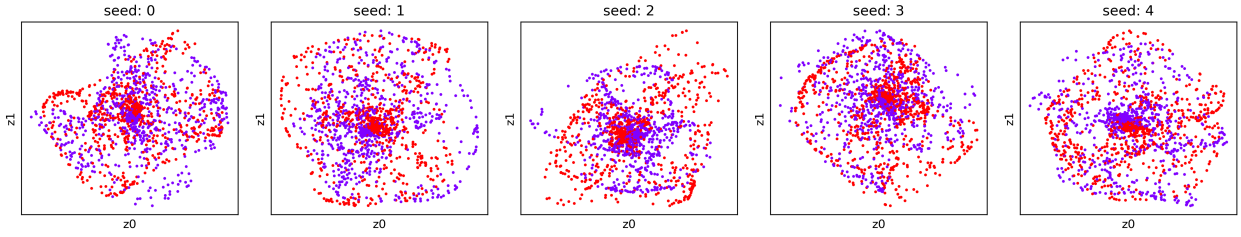
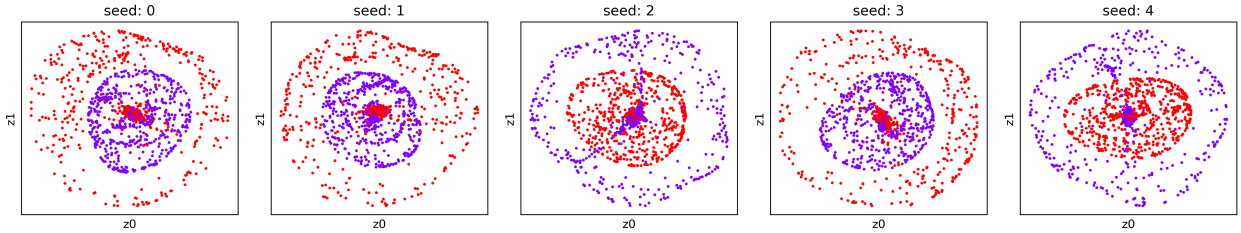


Figure 17: Exploration of latent space with negative label.

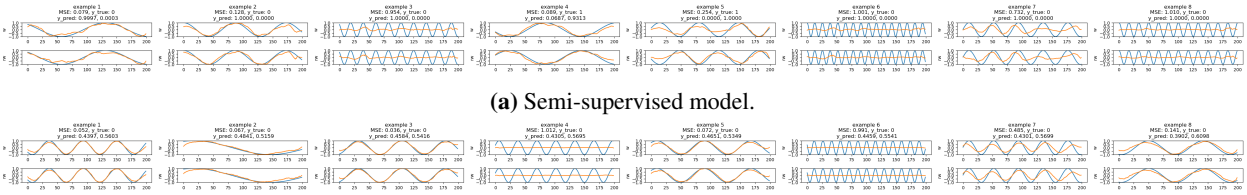


(a) Semi-supervised model.



(b) Unsupervised model.

Figure 18: Representations of sine-waves in latent space.



(a) Semi-supervised model.

(b) Unsupervised model.

Figure 19: Reconstruction of sine-waves.

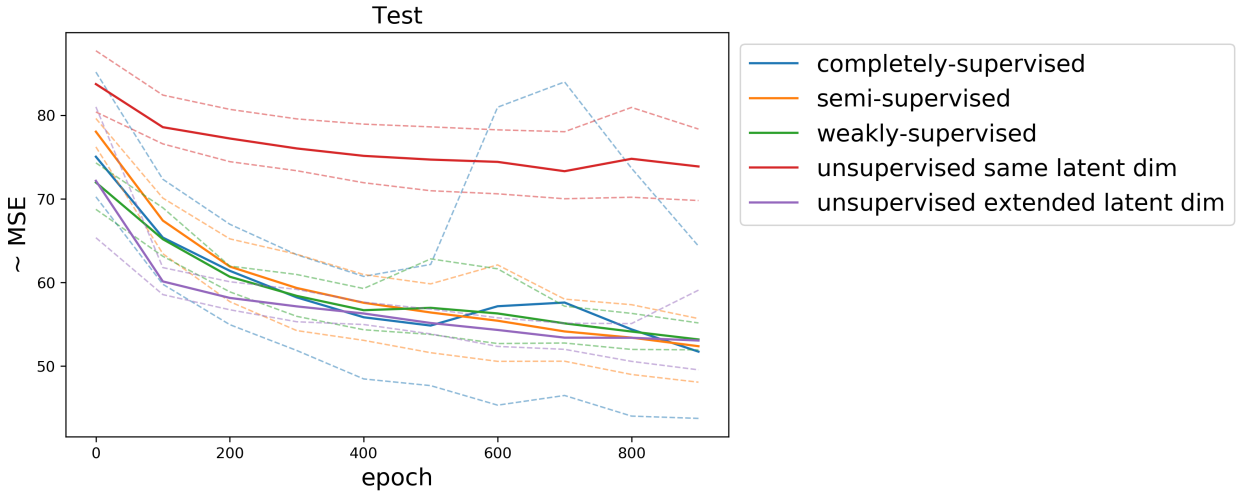


Figure 20: Comparison of different (semi-supervised) VAE models.

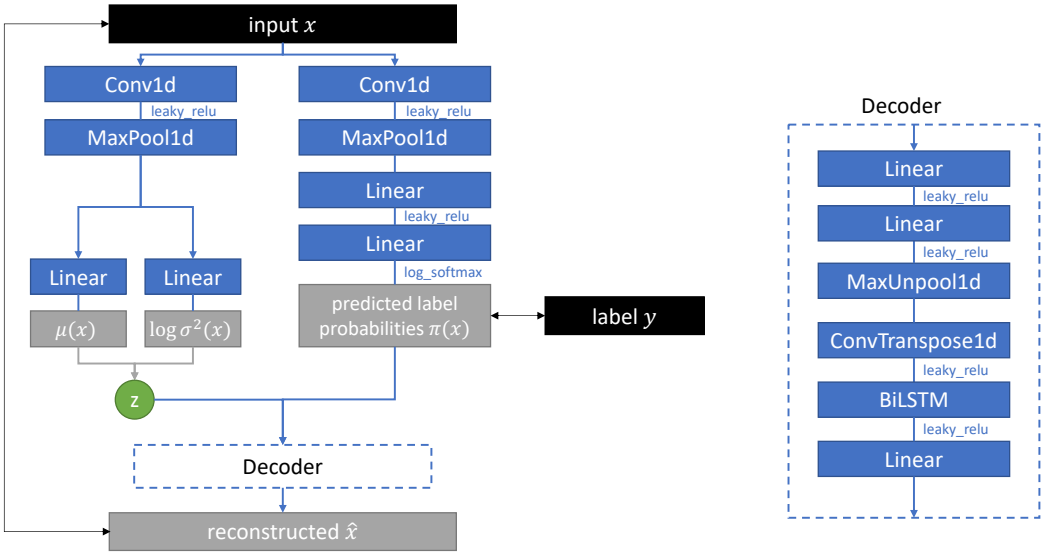
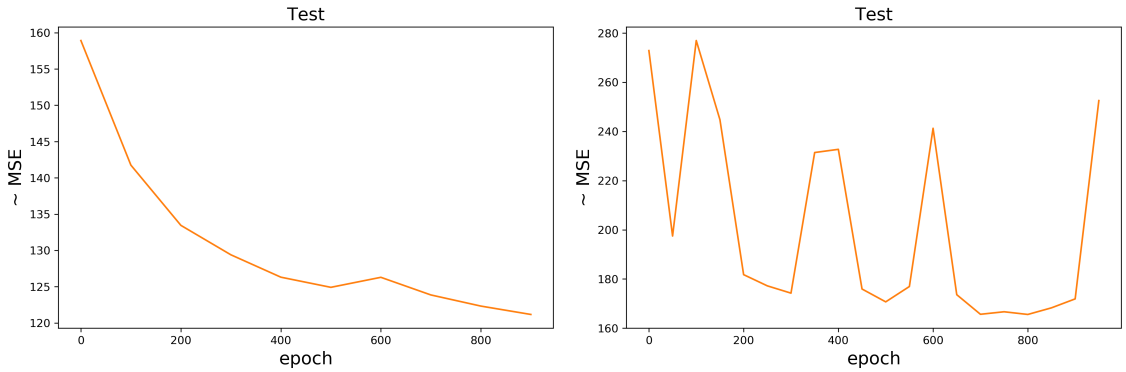
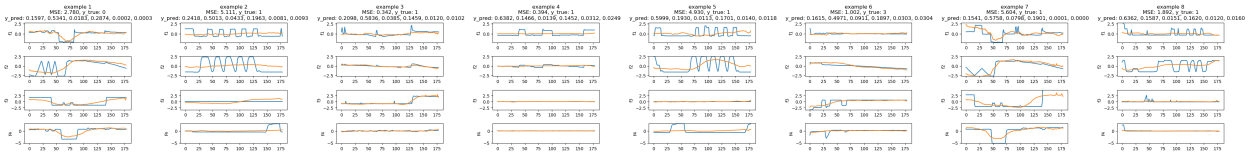


Figure 21: Semi-supervised VAE for time-series architecture.

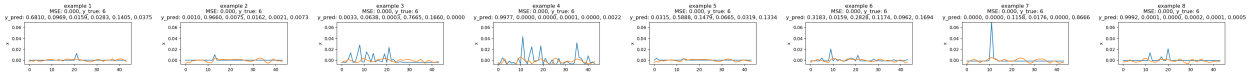


(a) MWD. **(b)** Inventory Demand.

Figure 22: Mean Squared Error on hold-out test samples from real datasets.



(a) MWD.



(b) Inventory Demand.

Figure 23: Semi-supervised reconstruction of time-series from real datasets.