Neural Decomposition of Time-Series Data

Anonymous Author(s)
Affiliation
Address
email

Abstract

We present a neural network technique for the analysis and extrapolation of time-series data called Neural Decomposition (ND). Units with a sinusoidal activation function are used to perform a Fourier-like decomposition of training samples into a sum of sinusoids, augmented by units with nonperiodic activation functions to capture linear trends and other nonperiodic components. We show how careful weight initialization can be combined with regularization to form a simple model that generalizes well. Our method generalizes effectively on a dataset of unemployment rates as reported by the U.S. Department of Labor Statistics, a time-series of monthly international airline passengers, and an unevenly sampled time-series of oxygen isotope measurements from a cave in north India. We find that ND outperforms popular time-series forecasting techniques including LSTM, echo state networks, ARIMA, SARIMA, and SVR with a radial basis function.

1 Introduction

The analysis and forecasting of time-series is a challenging problem that continues to be an active area of research. Predictive techniques have been presented for an array of problems, including weather [7], traffic flow [10], seizures [5], sales [3], and others [20, 21]. Because research in this area can be so widely applied, there is great interest in discovering more accurate forecasting methods.

One approach for analyzing time-series data is to interpret it as a signal and apply the Fourier transform to decompose the data into a sum of sinusoids [1]. Unfortunately, despite the well-established utility of the Fourier transform, it cannot be applied directly to time-series forecasting. Although the signal produced by the Fourier transform exactly reproduces the training samples, it also predicts that the same pattern of samples will repeat indefinitely. Another limitation of the Fourier transform is that it only uses periodic components, and thus cannot accurately model the nonperiodic aspects of a signal, such as a linear trend or nonlinear abnormality.

Another approach to forecasting time-series data is to use a model such as a neural network. Both feedfoward neural networks and recurrent neural networks have been applied to time-series analysis. Feedforward networks, such as Fourier neural networks that are initialized to compute the Fourier transform, have yielded promising results, but have also proven difficult to train [7]. Recurrent networks have tended to perform better at this task, with LSTM networks being among the most popular approaches [8]. Other models perform sinusoidal regression [22, 23], borrowing insights from harmonic analysis methods like the discrete Fourier transform. Regression and extrapolation-based approaches, however, have been largely abandoned in favor of the more successful recurrent models.

We claim that effective generalization can be achieved by regression and extrapolation using a model with two essential properties: (1) it must combine both periodic and nonperiodic components, and (2) it must be able to tune its components as well as the weights used to combine them. We present a neural network technique called Neural Decomposition (ND) that demonstrates this claim. ND decomposes a signal into a sum of constituent parts such that it can construct a signal that is useful

Submitted to 29th Conference on Neural Information Processing Systems (NIPS 2016). Do not distribute.

for extrapolating beyond the training samples. Furthermore, ND trains the components into which it decomposes the signal represented by training samples, enabling it to find a simpler set of constituent parts. In contrast to recurrent models like LSTM networks and harmonic analysis methods such as the discrete Fourier transform, ND does not require that samples be measured at regular intervals. Additionally, ND facilitates the inclusion of nonperiodic components, such as linear or sigmoidal components, to account for trends and nonlinear irregularities in a signal.

Models like ND have previously been proposed. Sinusoidal regression was studied as early as 1969 [22], and many papers have investigated related approaches, such as Fourier neural networks [17, 12, 7]. Our work differs from existing literature in at least two ways. First, we combine two simple insights (the necessity for periodic and nonperiodic components and the necessity for trainable components) in a simple, shallow feedforward neural network. Second, we demonstate that in some cases, this simple approach is able to outperform popular methods for time-series forecasting, such as LSTM. To our knowledge, no shallow network has been able to achieve this level of results before.

In Section 4, we demonstrate that the simple innovations of ND work together to produce significantly improved generalizing accuracy with several problems. We tested with a dataset of unemployment rates as reported by the U.S. Department of Labor Statistics, a time-series of monthly international airline passengers, and an unevenly sampled time-series of oxygen isotope measurements from a cave in north India. We compared against LSTM networks, echo state networks, ARIMA and SARIMA models, and SVR with a radial basis function. In each case, ND made better predictions than each of the other prediction techniques evaluated.

2 Related Work

2.1 Models for Time-Series Prediction

Many works have diligently surveyed the existing literature regarding forecasting techniques [2, 9, 4, 24]. Among the most popular statistical approaches are ARIMA models [24]. Seasonal ARIMA (SARIMA) is considered to be the state of the art "classical" time-series approach [10].

In the field of machine learning, three high-level classes of techniques are commonly used to forecast time-series data [7]. The first method, perhaps the most common approach, is to train a model to directly forecast future samples based on a sliding window of recently collected samples [6]. The second method is is to train a recurrent neural network [13]. Recurrent models, such as LSTM networks [8], have reported very good results for forecasting time-series. Our model falls into the third category of machine learning techniques: regression-based extrapolation. Our model is more closely related to a subclass of methods in this category, called Fourier neural networks (see Section 2.3). Models in the first two categories have already been well-studied, whereas extrapolation with sinusoidal neural networks remains relatively unexplored.

2.2 Harmonic Analysis

The harmonic analysis of a signal transforms a set of samples from the time domain to the frequency domain. This is useful in time-series prediction because the resulting frequencies can be used to reconstruct the original signal (interpolation) and to forecast values beyond the sampled time window (extrapolation). Harmonic analysis, also known as spectral analysis or spectral density estimation, has been well-studied for decades [14, 19].

Perhaps the most popular method of harmonic analysis is the discrete Fourier transform (DFT). The DFT maps a series of N complex numbers in the time domain to the frequency domain, and its inverse can be applied to map frequency domain values back to the time domain. For real-valued input, the inverse DFT can be sufficiently described as

$$x(t) = \sum_{k=0}^{N/2} R_k \cdot cos(\frac{2\pi k}{N} t) - I_k \cdot sin(\frac{2\pi k}{N} t). \tag{1}$$

Equation 1 is useful as a continuous representation of the real-valued discrete input. Because it perfectly passes through the input samples, one might naively expect this function to be a good basis

for generalization. However, the DFT cannot effectively model the nonperiodic components of a signal, nor can it form a simple model for series that are not periodic at N.

Because of these limitations, other approaches to the harmonic analysis of time-series have been proposed. Some of these other approaches perform sinusoidal regression to determine frequencies that better represent the periodicity of the sampled signal [22, 23]. However, these approaches have largely been abandoned in favor of the fast Fourier transform, which allows the DFT to be calculated in $log(n)$ time. Like these less popular approaches, our approach uses regression to find better frequencies. Our results show that this antiquiated regression-based approach, combined with nonlinear components, is able to outperform state of the art methods for time-series prediction.

2.3 Fourier Neural Networks

Use of the Fourier transform in neural networks has already been explored in various contexts. The term *Fourier neural network* has been used to refer to neural networks that use a Fourier-like neuron [17], that use the Fourier transform of some data as input [12], or that use the Fourier transform of some data as weights [7]. Our work is not technically a Fourier neural network, but of these three types, our approach most closely resembles the third.

Silvescu provided a model for a Fourier-like activation function for neurons in neural networks [17]. His model utilizes every unit to form DFT-like output for its inputs. He notes that by using gradient descent to train sinusoid frequencies, the network is able to learn "exact frequency information" as opposed to the "statistical information" provided by the DFT. Our approach also trains the frequencies of neurons with a sinusoidal activation function.

Gashler and Ashmore presented a technique that used the fast Fourier transform (FFT) to approximate the DFT, then used the obtained values to initialize the sinusoid weights of a neural network that mixed sinusoidal, linear, and softplus activation functions [7]. Because this initialization used sinusoid units to model nonperiodic components of the data, their model was designed to heavily regularize sinusoid weights so that as the network was trained, it gave preference to weights associated with nonperiodic units and shifted the weights from the sinusoid units to the linear and softplus units. Use of the FFT required their input size to be a power of two, and their trained models were slightly out of phase with their validation data. However, they were able to generalize well for certain problems. Our approach is similar, except that we do not use the Fourier transform to initialize any weights.

3 Approach

In this section, we describe Neural Decomposition (ND), a neural network technique for the analysis and extrapolation of time-series data.

3.1 High-Level Description

We use a DFT-like model with two simple but important innovations. First, we allow sinusoid frequencies to be trained. Second, we augment the sinusoids with a nonperiodic function. The use of sinusoids allows our model to fit to periodic data, the ability to train the frequencies allows our model to learn the true period of a signal, and the augmentation function enables our model to forecast time-series that are made up of both periodic and nonperiodic components.

Our model is defined as follows. Let each a_k represent an amplitude, each w_k represent a frequency, and each ϕ_k represent a phase shift. Let t refer to time, and let $g(t)$ be an augmentation function that represents the nonperiodic components of the signal. Our model is defined as

$$x(t) = \sum_{k=1}^{N} \big(a_k \cdot sin(w_k t + \phi_k)\big) + g(t). \tag{2}$$

Note that the lower index of the sum has changed from $k = 0$ in the DFT to $k = 1$ in our model. This is because ND can account for bias in the augmentation function $g(t)$, so the 0 frequency is not necessary. Therefore, only N sinusoids are required rather than $N + 2$.

3.2 Topology

We use a feedforward artificial neural network as the basis of our model. For an input of size N, the neural network is initialized with two layers with sizes $1 \rightarrow m$ and $m \rightarrow 1$, where $m = N + |g(t)|$ and $|g(t)|$ denotes the number of nodes required by $g(t)$. The first N nodes in the hidden layer have the sinusoid activation function, $sin(t)$, and the rest of the nodes in the hidden layer have other activation functions to compute $g(t)$.

The augmentation function $g(t)$ can be made up of any number of nodes with one or more activation functions. For example, it could be made up of linear units for learning trends and sigmoidal units to fit nonlinear irregularities. In our experiments, we used a combination of linear, softplus, and sigmoidal nodes for $g(t)$. The network tended to only use a single linear node, which may suggest that the primary benefit of the augmentation function is that it can model linear trends in the data. Softplus and sigmoidal units tended to be used very little or not at all by the network in the problems we tested, but intuitively it seems that nonlinear activation functions could be useful in some cases.

3.3 Weight Initialization

The weights of the neural network are initialized as follows. Let each of the N sinusoid nodes in the hidden layer, indexed as k for $0 \leq k < N/2$, have a weight w_k and bias ϕ_k. Each w_k represents a frequency and is initialized to $2\pi\lfloor k/2 \rfloor$. Each ϕ_k represents a phase shift. For each even value of k, ϕ_k is set to $\pi/2$ to transform $sin(t + \phi_k)$ to $cos(t)$. For each odd value of k, ϕ_k is set to π to transform $sin(t + \phi_k)$ to $-sin(t)$. A careful comparison of these initialized weights with Equation 1 shows that these are identical to the frequencies and phase shifts used by the DFT, except for a missing $1/N$ term in each frequency, which is absorbed in the input preprocessing step (see Section 3.4).

All weights feeding into the output unit are set to small random values. At the beginning of training, therefore, the model will predict something like a flat line centered at zero. As training progresses, the neural network will learn how to combine the hidden layer units to fit the training data.

Weights in the hidden layer associated with the augmentation function are initialized to approximate the identity function. For example, in $g(t) = wt + b$, w is randomly perturbed from 1 and b is randomly perturbed near 0. Because the output layer will learn how to use each unit in the hidden layer, it is important that each unit be initialized in this way.

3.4 Input Preprocessing

Before training begins, we preprocess the input data to facilitate learning and prevent the model from falling into a local optimum. First, we normalize the time associated with each sample so that the training data lies between 0 (inclusive) and 1 (exclusive) on the time axis. If there is no explicit time, equally spaced values between 0 and 1 are assigned to each sample in order. Predicted data points will have a time value greater than or equal to 1 by this new scale. Second, we normalize the values of each input sample so that all training data is between 0 and 10 on the y axis.

This preprocessing step serves two purposes. First, it absorbs the $1/N$ term in the frequencies by transforming t into t/N, which is why we were able to omit the $1/N$ term from our frequencies in the weight initialization step. Second, it ensures that the data is appropriately scaled so that the neural network can learn efficiently. If the data is scaled too large, training will be slow and susceptible to local optima. If the data is scaled too small, on the other hand, the learning rate of the machine will cause training to diverge and only use linear units and low frequency sinusoids.

In some cases, it is also appropriate to pass the input data through a filter. For example, financial time-series data is commonly passed through a logarithmic filter before being presented for training, and outputs from the model are then exponentiated to obtain predictions. We use this filtering method in two of our experiments in Section 4.

3.5 Regularization

Prior to each sample presentation, we apply regularization on the output layer of the neural network. Even though we do not initialize sinusoid amplitudes using the DFT, the network is quickly able to learn how to use the initialized frequencies to closely fit the input samples. Without regularizing the output layer, training halts as soon as the model fits the input samples, because the measurable

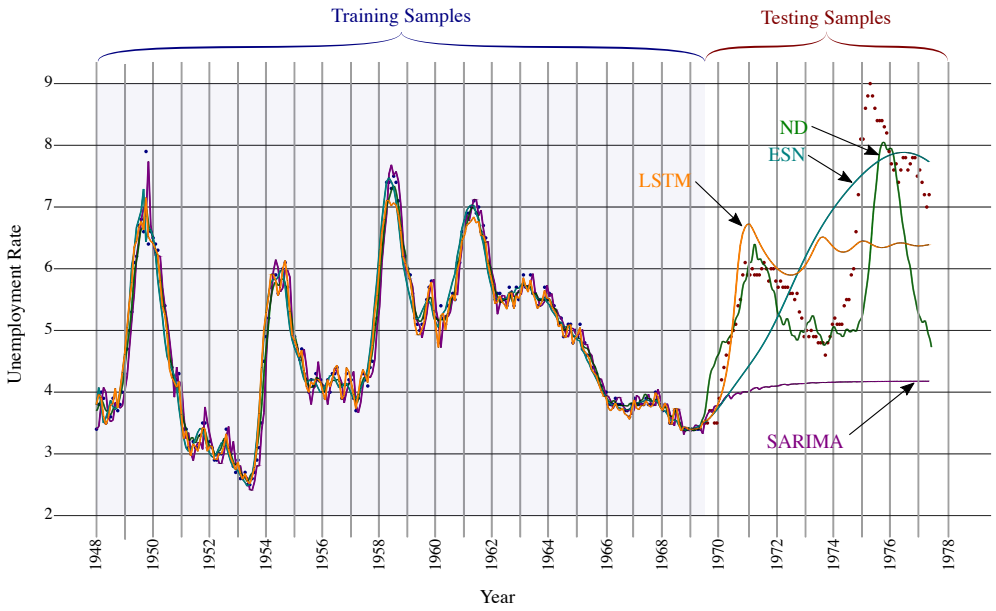

Figure 1: A comparison of the four best predictive models on the monthly unemployment rate in the US. Blue points represent training samples from January 1948 to June 1969 and red points represent testing samples from July 1969 to December 1977. Only ND, shown in green, successfully predicted both surges in unemployment that followed the training samples.

error is near zero. By relaxing the learned weights, regularization allows our model to redistribute its weights over time. We find that regularization amount is especially important; too much prevented our model from learning, but too little caused our model to fall into local optima. In our experiments, setting the regularization term to 10^{-2} avoided both of these potential pitfalls.

Another important function of regularization in ND is to promote sparsity in the network, so that the redistribution of weights by stochastic gradient descent produces as simple a model as the input samples allow. We use L^1 regularization for this reason. Usually, the trained model does not require all N sinusoid nodes in order to generalize well, and this type of regularization enables the network to automatically discard unnecessary nodes by driving their amplitudes to zero.

It is worth noting that we only apply regularization to the output layer of the neural network. Any regularization that might occur in the hidden layer would adjust sinusoid frequencies before the output layer could learn sinusoid amplitudes. By allowing weights in the hidden layer to change without regularization, the network has the capacity to adjust frequencies but is not required to do so.

4 Validation

In this section, we report results from three of our experiments that validate the effectiveness of Neural Decomposition. In each of these experiments, we used an ND model with an augmentation function made up of ten linear units, ten softplus units, and ten sigmoidal units. It is worth noting that $g(t)$ is under no constraint to consist only of these units; it could include other activation functions or only contain a single linear node. We use a regularization term of 10^{-2} and a learning rate of 10^{-3} in every experiment to demonstrate the robustness of our approach. We did not tune the meta-parameters of ND for each experiment, but we did carefully tune the meta-parameters of competing models for each experiment using a grid search.

In our experiments, we compare ND with LSTM, ESN, ARIMA, SARIMA, and SVR. We used PyBrain's implementation of LSTM networks [16] with one input neuron, one output neuron, and one hidden layer. We implemented a grid-search to find the best hidden layer size for the LSTM network for each problem and used PyBrain's RPROP- algorithm to train the network. We used Lukoševičius' implementation of ESN [11] and implemented a grid-search to find the best parameters for each

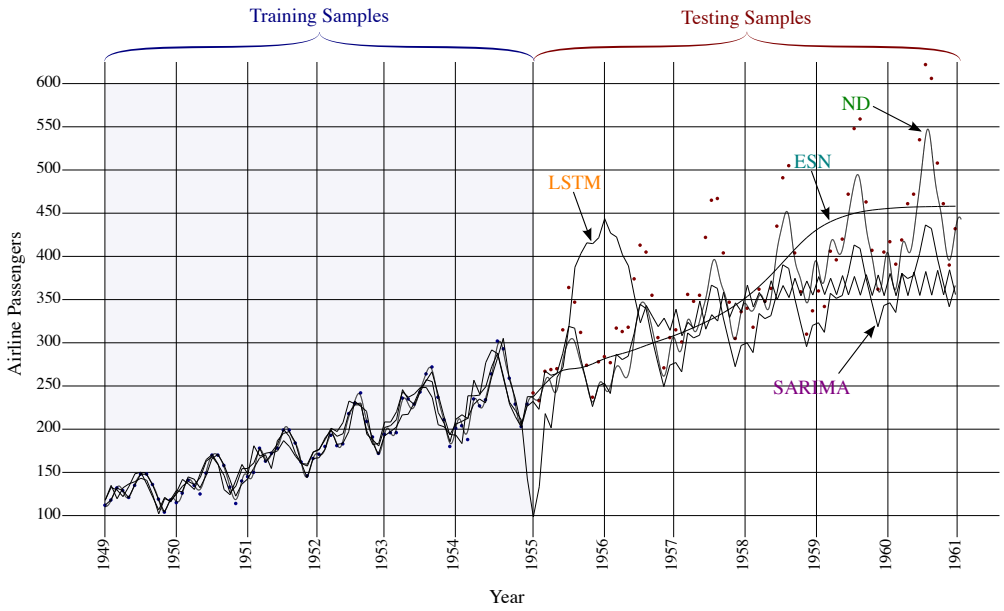

Figure 2: A comparison of the four best predictive models on monthly totals of international airline passengers from January 1949 to December 1960 [2]. Blue points represent the 72 training samples from January 1949 to December 1954 and red points represent the 72 testing samples from January 1955 to December 1960. ND, shown in green, learns the trend, shape, and growth better than the other compared models.

problem. We used the R language implementation for ARIMA, SARIMA, and SVR [15]. For the ARIMA models, we used a variation of the `auto.arima` method that performs a grid-search to find the best parameters for each problem. For SVR, we used the `tune.svm` method, which also performs a grid-search for each problem. Although these methods select the best models based on the amount of error calculated using the training samples, the grid-search is a very slow process. With ND, no problem-specific parameter tuning was performed.

In each figure, the blue points in the shaded region represent training samples and the red points represent withheld testing samples. The curves on the graph represent the predictions made by the four models that made the most accurate predictions (only two models are shown in the fourth experiment because only two models could be applied to an irregularly sampled time-series). The actual error for each model's prediction is reported for all experiments and all models in Table 1.

The LSTM network tended to fall into local optima, and was thus very sensitive to the random seed. Running the same experiment with LSTM using a different random seed yielded very different results. In each experiment, therefore, we tried the LSTM model 100 times for each topology tested in our grid-search and selected the result with the highest accuracy to present for comparison with ND. Conversely, ND consistently made approximately identical predictions when run multiple times, regardless of the random seed.

In our first experiment, we demonstrate the effectiveness of ND on real-world data compared to widely used techniques in time-series analysis and forecasting. We trained our model on the unemployment rate from 1948 to 1969 as reported by the U.S. Bureau of Labor Statistics, and predicted the unemployment rate from 1969 to 1977. These results are shown in Figure 1. Blue points on the left represent the 258 training samples from January 1948 to June 1969, and red points on the right represent the 96 testing samples from July 1969 to December 1977. The four curves represent predictions made by ND (green), LSTM (orange), ESN (cyan), and SARIMA (magenta); ARIMA and SVR yielded poorer predictions and are therefore omitted from the figure. Grid-search found ARIMA(3,1,2) and ARIMA(1,1,2)(1,0,1)[12] for the ARIMA and SARIMA models, respectively. ARIMA, not shown, did not predict the significant rise in unemployment. SARIMA, shown in magenta, did correctly predict a rise in unemployment, but underestimated its magnitude, and did not predict the shape of the data well. SVR, not shown, correctly predicted that unemployment would

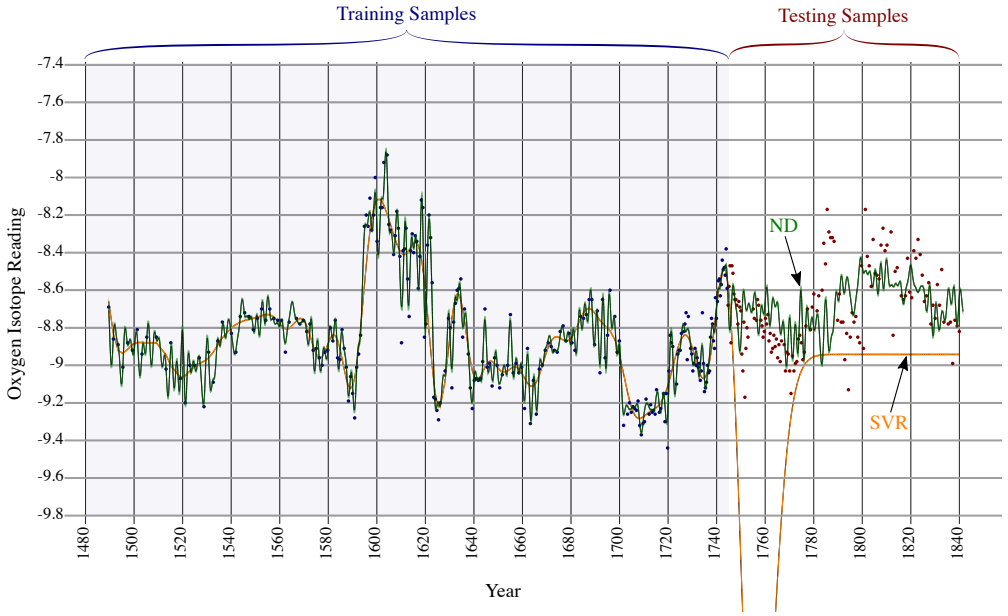

Figure 3: A comparison of two predictive models on a series of oxygen isotope readings in speleothems in India from 1489 AD to 1839 AD [18]. Blue points represent the 250 training samples from July 1489 to April 1744 and red points represent the 132 testing samples from August 1744 to December 1839. Only ND and SVR could be applied to this irregularly sampled time-series. ND, shown in green, is the only model that captures the general shape of the testing samples.

rise, then fall again. However, it also underestimated the magnitude. ESN, shown in cyan, predicted a reasonable mean value for the general increase in unemployment, but failed to capture the dynamics of the actual data. The LSTM network (shown in orange), with a hidden layer of size 16 found by grid-search, predicted the first peak in the data, but leveled off to predict only the mean.

Results with Neural Decomposition (ND) are shown in green. ND successfully predicted both the depth and approximate shape of the surge in unemployment. Furthermore, it correctly anticipated another surge in unemployment that followed. ND did a visibly better job of predicting the nonlinear trend much farther into the future.

Our next experiment demonstrates the versatility of Neural Decomposition in its application to another real-world dataset: monthly totals of international airline passengers as reported by Chatfield [2]. We used the first six years of data (72 samples) from January 1949 to December 1954 as training data, and the remaining six years of data (72 samples) from January 1955 to December 1960 as testing data. The training data was preprocessed through a $log(x)$ filter and the outputs were exponentiated to obtain the final predictions. As in the first experiment, we compared our model with LSTM, ESN, ARIMA, SARIMA, and SVR. The predictions of the four most accurate models (ND, LSTM, ESN, and SARIMA) are shown in Figure 2; ARIMA and SVR yielded poorer predictions and are therefore omitted from the figure. SVR, not shown, predicted a flat line after the first few time steps and generalized the worst out of the four predictive models. The ARIMA model found by grid-search was ARIMA(2,1,3). ARIMA, not shown, was able to learn the trend, but failed to capture any of the dynamics of the signal. Grid-search found ARIMA(1,0,0)(1,1,0)[12] for the SARIMA model. Both SARIMA (shown in magenta) and ND (shown in green) were able to accurately predict the shape of the future signal, but ND performed better. Unlike SARIMA, ND learned that the periodic component gets bigger over time. ESN, shown in cyan, performed similarly to the ARIMA model, only predicting the trend and failing to capture seasonal variations. The LSTM network (shown in orange), with a hidden layer of size 64 found by grid-search, also failed to capture any meaningful seasonality in the training data. Instead, LSTM immediately predicted a valley and a peak that did not actually occur, followed by a poor estimation of the mean.

Our third experiment demonstrates that ND can be used on irregularly sampled time-series. We used a series of oxygen isotope readings in speleothems in a cave in India from 1489 AD to 1839 AD as

Table 1: MAPE of all models on the validation problems. Smallest error for each is shown in **bold**.

Model	Labor	Airline	Isotope
ARIMA	39.42%	12.34%	N/A
SARIMA	29.69%	13.33%	N/A
SVR	25.14%	47.04%	8.50%
ESN	15.73%	12.05%	N/A
LSTM	14.63%	18.95%	N/A
ND	**10.89%**	**9.52%**	**1.89%**

reported by Sinha et. al [18]. Because the time intervals between adjacent samples is not constant (the interval is about 1.5 years on average, but fluctuates between 0.5 and 2.0 years), only ND and SVR models could be applied. ARIMA, SARIMA, ESN, and LSTM cannot be applied to irregular time-series because they assume a constant time interval between adjacent samples; these five models are therefore not included in this experiment. Figure 3 shows the predictions of ND and SVR. Blue points on the left represent the 250 training samples from July 1489 to April 1744, and red points on the right represent the 132 testing samples from August 1744 to December 1839. SVR, shown in orange, predicted a steep drop in value that does not exist in the testing data. ND, shown in green, accurately predicted the general shape of the testing data.

Table 1 presents an empirical evaluation of each model for the three real-world experiments. We use the mean absolute percent error (MAPE) as our error metric for comparisons [10]. For a set of n predictions $x(t)$ and a set of n samples x_t, MAPE is defined as $\frac{1}{n}\sum_{t=1}^{n}\left|\frac{x_t - x(t)}{x_t}\right|$. Using MAPE, we compare Neural Decomposition to ARIMA, SARIMA, SVR, ESN, and LSTM. We found that on the unemployment rate problem (Figure 1), ND yielded the best model, followed by LSTM and ESN. On the airline problem (Figure 2), ND performed significantly better than all of the other approaches. On the oxygen isotope problem (Figure 3), ND outperformed SVR, which was the only other model that could be applied to the irregular time-series. Table 1 presents the results of our experiments. In each problem, the accuracy of the best algorithm is shown in bold.

5 Conclusion

We presented Neural Decomposition, a neural network technique for time-series forecasting. Our method decomposes a set of training samples into a sum of sinusoids augmented with additional components to enable our model to generalize and extrapolate beyond the input set. Each component of the resulting signal is trained so that ND can find a simpler set of constituent signals. ND uses careful initialization, input preprocessing, and regularization to facilitate the training process. We showed results that demonstrate that our approach is superior to popular techniques LSTM, ESN, ARIMA, SARIMA, and SVR in some cases, including the US unemployment rate, monthly airline passengers, and an unevenly sampled time-series of oxygen isotopes. We predict that ND will similarly perform well on a number of other problems.

This work makes the two following contributions to the current knowledge. First, it identifies two properties necessary for a regression and extrapolation-based model to be effective: trainable components and nonperiodic augmentation. Second, it demonstrates that this antiquated regression-based approach to time-series analysis is still useful and can outperform state of the art techniques for some problems.

References

[1] P. Bloomfield. *Fourier analysis of time series: an introduction*. John Wiley & Sons, 2004.

[2] C. Chatfield. *The Analysis of Time Series: An Introduction*. Chapman and Hall/CRC, 6 edition, July 2003.

[3] T.-M. Choi, Y. Yu, and K.-F. Au. A hybrid sarima wavelet transform method for sales forecasting. *Decision Support Systems*, 51(1):130–140, 2011.

[4] G. Dorffner. Neural networks for time series processing. In *Neural Network World*. Citeseer, 1996.

[5] C. E. Elger and K. Lehnertz. Seizure prediction by non-linear time series analysis of brain electrical activity. *European Journal of Neuroscience*, 10(2):786–789, February 1998.

[6] R. J. Frank, N. Davey, and S. P. Hunt. Time series prediction and neural networks. *Journal of Intelligent and Robotic Systems*, 31(1-3):91–103, 2001.

[7] M. S. Gashler and S. C. Ashmore. Training deep fourier neural networks to fit time-series data. In *Intelligent Computing in Bioinformatics - 10th International Conference, ICIC 2014, Taiyuan, China, August 3-6, 2014. Proceedings*, pages 44–55, 2014.

[8] F. A. Gers, J. Schmidhuber, and F. Cummins. Learning to forget: Continual prediction with lstm. *Neural computation*, 12(10):2451–2471, 2000.

[9] I. Kaastra and M. Boyd. Designing a neural network for forecasting financial and economic time series. *Neurocomputing*, 10(3):215–236, 1996.

[10] M. Lippi, M. Bertini, and P. Frasconi. Short-term traffic flow forecasting: An experimental comparison of time-series analysis and supervised learning. In *IEEE Transactions on Intelligent Transportation Systems*, volume 14, pages 871–882, March 2013.

[11] M. Lukoševičius. *A practical guide to applying echo state networks*, volume 7700 of *Lecture Notes in Computer Science*, pages 659–686. Springer Berlin Heidelberg, 2 edition, 2012.

[12] K. Minami, H. Nakajima, and T. Toyoshima. Real-time discrimination of ventricular tachyarrhythmia with fourier-transform neural network. *IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING*, 46(2), February 1999.

[13] O. Nerrand, P. Roussel-Ragot, D. Urbani, L. Personnaz, and G. Dreyfus. Training recurrent neural networks: Why and how ? an illustration in dynamical process modeling., 1994.

[14] B. G. Quinn. Estimating the number of terms in a sinusoidal regression. *Journal of time series analysis*, 10(1):71–75, 1989.

[15] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[16] T. Schaul, J. Bayer, D. Wierstra, Y. Sun, M. Felder, F. Sehnke, T. Rückstieß, and J. Schmidhuber. PyBrain. *Journal of Machine Learning Research*, 11:743–746, 2010.

[17] A. Silvescu. Fourier neural networks. In *Neural Networks, 1999. IJCNN '99. International Joint Conference on*, volume 1, pages 488–491, 1999.

[18] A. Sinha, G. Kathayat, H. Cheng, S. F. M. Breitenbach, M. Berkelhammer, M. Mudelsee, J. Biswas, and R. L. Edwards. Trends and oscillations in the indian summer monsoon rainfall over the last two millennia. *Nat Commun*, 6, 02 2015.

[19] E. M. Stein and T. S. Murphy. *Harmonic analysis: real-variable methods, orthogonality, and oscillatory integrals*, volume 3. Princeton University Press, 1993.

[20] F. E. Tay and L. Cao. Application of support vector machines in financial time series forecasting. *Omega*, 29(4):309–317, August 2001.

[21] J. W. Taylor, P. E. McSharry, and R. Buizza. Wind power density forecasting using ensemble predictions and time series models. *Energy Conversion, IEEE Transactions on*, 24(3):775–782, September 2009.

[22] P. Vanicek. Approximate spectral analysis by least-squares fit. *Astrophysics and Space Science*, 4(4):387–391, 1969.

[23] Y. Yokouchi. A strong source of methyl chloride to the atmosphere from tropical coastal land. *Nature*, 2000.

[24] G. P. Zhang. Time series forecasting using a hybrid arima and neural network model. *Neurocomputing*, 50:159–175, January 2003.