Time Series Forecasting for Solar Irradiance Data Using Matrix Estimation Methods

Esha Srivastava
Department of Mathematics, Northeastern University

August 22, 2023

Abstract

Solar irradiance data is crucial for various applications, including renewable energy forecasting, climate research, and environmental monitoring. In this project, the aim is to present a forecasting algorithm that leverages matrix estimation methods to impute and forecast time series data. This approach operates in a model and noise agnostic setting, making it applicable solar irradiance forecasting without relying on specific assumptions about the underlying dynamics or noise distribution.

Keywords: Time Series, Matrix Estimation, Solar Irradiance

1 Introduction

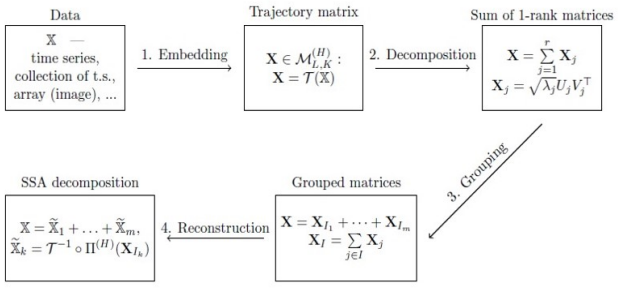
Time series data plays a pivotal role in the analysis of energy systems, encompassing everything from wind and weather predictions to the exploration of mining and measurement of solar radiance. This kind of data, which evolves over time, is widely prevalent. However, despite its immense potential, time series data is not immune to challenges. The concept of noise and missing data introduces complexities that can substantially affect the reliability and accuracy of analysis outcomes.

Noise, arising from factors such as measurement errors or inherent variability, can distort patterns and trends within the data. Similarly, the presence of missing data points can introduce bias and hinder the formulation of comprehensive insights. These challenges, if not appropriately addressed, can mislead decision-making processes and compromise the quality of outcomes. Therefore, the development of robust methodologies capable of handling noise and missing data becomes paramount in ensuring the integrity of time series analysis.

Among the core challenges faced in time series analysis are interpolation and extrapolation. Interpolation involves the task of accurately estimating data points within the observed range, enhancing the granularity of insights. On the other hand, extrapolation delves into the realm of predicting future trends beyond the scope of available data. These challenges are not constrained by specific domains or data types; rather, they persistently surface across various applications. The accurate handling of these challenges contributes not only to the precision of analysis but also to the reliability of predictive models.

In light of these complexities and challenges, this work explores novel approaches to address the issues inherent in time series analysis. By leveraging advancements in machine learning and data science, the aim is to provide effective solutions that bolster the accuracy, reliability, and utility of time series data in diverse fields. The subsequent sections explore the methodologies, algo-

rithm, properties and experiment on solar irradiance dataset, shedding light on the intricacies of time series data analysis and its overarching impact.



(Singular Spectrum Analysis [3])

2 Methodology

The idea behind the approach here is to think about time series as a matrix. Consider a discrete-time setting with $t \in \mathbb{Z}$ representing the time index and $f : \mathbb{Z} \to \mathbb{R}^1$ representing the latent discrete-time time series of interest. The time series data, $X(t)$ for $t \in [T]$, is transformed where $T = \{0, 1, \ldots\}$, [1] into a page matrix by segmenting contiguous segments of size $L > 1$ into non-overlapping columns.

This transformation offers insights into the underlying patterns and relationships within the time series data. The page matrix is a creative solution to the challenges posed by the Hankel matrix's unique block structure. It stems from Markov parameters and provides a simple yet effective way of organizing data.[2]

Previous Methods for Matrix Estimation:

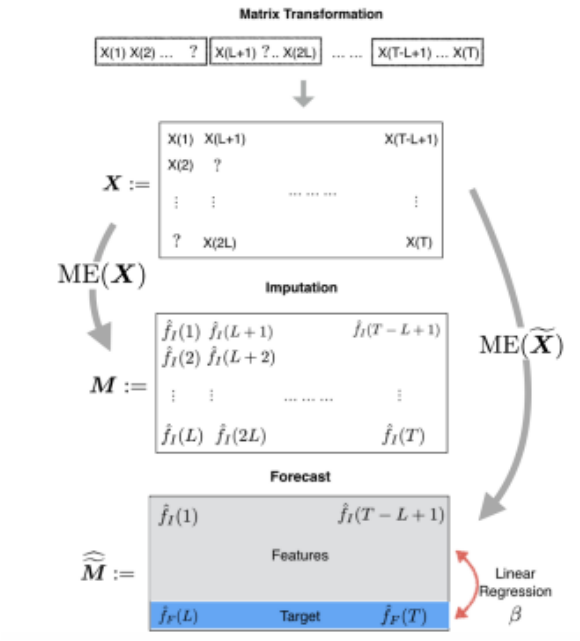
Representing time series as a matrix is not new. In Singular Spectrum Analysis (SSA), the core steps are as follows: (1) create a Hankel matrix from the time series data; (2) perform a Singular Value Decomposition (SVD) of it; (3) group the singular values based on the user's belief of the model that generated the process; (4) perform diagonal averaging for the "Hankelization" of the grouped rank-1 matrices outputted from the SVD to create a set of time series; (5) learn a linear model for each "Hankelized" time series for the purpose of forecasting.[1]

3 Algorithm

Matrix Transformation First, the noisy time series $X(t)$ (with "?" indicating missing data) is transformed into a page matrix X with non-overlapping entries.

Imputation: $ME(X) \quad X \to M$
A matrix estimation (ME) algorithm is applied with an input X to obtain the estimates $\hat{f}_I(t)$ for the de-noised and filled-in entries.

Forecasting: $ME(\tilde{X}) \quad X \to \tilde{X} \to \hat{M}_F$
The ME is applied to $\tilde{X}$ (i.e., X excluding the last row), and then is fit with a linear model β (Noisy regression) between the last row and all other rows to obtain the forecast estimates $\hat{f}_F(t)$.



(Illustration taken from [1] to show the steps of the algorithm)

Singular Spectrum Analysis(SSA) vs. Matrix Estimation(ME)Algorithm

In SSA, there are the following properties:

- A Hankel matrix, which has repeated entries
- A heavily reliance on Singular Value Decomposition (SVD)
- An assumption that the underlying data is fully observed and noiseless

In the ME Algorithm, there are the following properties:

- A page matrix, which has non-overlapping entries
- A utilization of general matrix estimation procedures, whereas SVD methods represent one specific procedural choice
- The use of noisy regression, which allows for corrupted and missing entries

At the heart of the algorithm lies matrix estimation. This is explored more in depth in the next section.

4 Problem Setup

Consider an $m \times n$ matrix M of interest. Suppose there is a random subset of the entries of a noisy signal matrix X, such that $E[X] = M$. For each $i \in [m]$ and $j \in [n]$, the (i, j)-th entry X_{ij} is a random variable that is observed with probability $p \in (0, 1]$ and is missing with probability $1 - p$, independently of all other entries. Given X, the goal is to produce an estimator $\hat{M}$ that is "close" to M. The following two metrics are used to quantify this estimation error:

$$\text{Mean Squared Error (MSE)}(\hat{M}, M) := E\left[\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(\hat{M} - M_{ij})^2\right]$$

$$\text{Max Row Sum Error (MRSE)}(\hat{M}, M) := E\left[\frac{1}{\sqrt{n}}\max\left(\sum_{j=1}^{n}(\hat{M} - M_{ij})^2\right)^{1/2}\right]$$

A matrix estimation algorithm [1], denoted as $ME : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$, takes as input a noisy matrix X and outputs an estimator M.

4.1 Properties for Matrix Estimation

Required Properties for matrix estimation algorithm ME(·)[1] to achieve theoretical guarantees:

Property 4.1.1. Let ME satisfy the following: Define $Y = [Y_{ij}]$ where $Y_{ij} = X_{ij}$ if X_{ij} is observed, and $Y_{ij} = 0$ otherwise. Then, for all $p \geq \max(m, n)^{-1+\zeta}$ and some $\zeta \in (0, 1)$, the produced estimator $M_b = ME(X)$ satisfies

$$\|\hat{p}M_b - pM\|_F^2 \leq \frac{1}{mn}C1\|Y - pM\|\|pM\|_*,$$

where $\hat{p}$ denotes the proportion of observed entries in X and $C1$ is a universal constant.

Property 4.1.2 Let ME satisfy the following: For all $p \geq p^*(m, n)$, the produced estimator $M_b = ME(X)$ satisfies

$$MRSE(M_b, M) \leq \delta_3(m, n),$$

where $\lim_{m,n\to\infty} \delta_3(m,n) = 0$.

4.2 Properties for Time Series

Required properties for the matrices $X^{(k)}$ and $M^{(k)}$ to identify the time series models[1] f:

4.2.1 Property 1: Imputable

For each $i \in [L]$ and $j \in [N]$:

1. $X_{ij}^{(1)}$ are independent sub-gaussian random variables satisfying $E[X_{ij}^{(1)}] = M_{ij}^{(1)}$ and $\|X_{ij}^{(1)}\|_{\psi_2} \leq \sigma$. $X_{ij}^{(1)}$ is observed with probability $p \in (0,1]$, independent of other entries.

2. There exists a matrix $M_r^{(1)}$ of rank r such that for $\delta_1 \geq 0$, $\|M^{(1)} - M_r^{(1)}\|_{\max} \leq \delta_1$.

4.2.2 Property 2: Forecastable

For all $k \in [L]$, let matrices $X^{(k)}$ and $M^{(k)}$ satisfy the following:

1. For each $i \in [L]$ and $j \in [N]$

   (a) $X_{ij}^{(k)} = M_{ij}^{(k)} + \epsilon_{ij}$, where ϵ_{ij}^k are independent sub-Gaussian random variables satisfying $E[\epsilon_{ij}^k] = 0$ and $\text{Var}(\epsilon_{ij}^k) \leq \sigma^2$.

   (b) $X_{ij}^{(k)}$ is observed with probability $p \in (0,1]$, independent of other entries.

2. There exists $\beta^{(k)} \in \mathbb{R}^{(L-1)}$ with $\|\beta^{*(k)}\|_1 \leq C_\beta$ for some constant $C_\beta > 0$ and $\delta_2 \geq 0$ such that

$$\|M_L^{(k)} - (\widetilde{M}^{(k)})^T \beta^{*(k)}\|_2 \leq \delta_2.$$

So far, only the important properties used to implement the algorithm have been discussed. For a more detailed explanation about the theorems and properties, please refer [1]

5 Experiment on Solar Irradiance Data

Solar irradiance, a cornerstone of renewable energy research, holds profound significance in optimizing solar energy systems. The temporal nature of irradiance data necessitates specialized time series analysis to unveil hidden patterns and trends. By taming the noise and filling in the gaps, better predictions about future solar irradiance can be made.

5.1 Data Overview

The dataset utilized in this study was provided by National Renewable Energy Laboratory (NREL)[4] Spanning the period from 2019 to 2021, this data set encapsulates a comprehensive record of solar irradiance data. Focusing the investigation on the geographical expanse of Massachusetts, there is a specific column of paramount importance: the Direct Normal Irradiance (DNI). This column is crucial because it offers insights into the temporal dynamics of solar energy availability in the specified region.

index	Year	Month	Day	Hour	Minute	DHI	Temperature	Clearsky DHI	Clearsky DNI	Clearsky GHI	Cloud Type	Dew Point	DNI	Fill Flag	GHI	Ozone	Relative Humidity	Solar Zenith Angle	Surface Albedo	Pressure
0	2019	1	1	0	0	0	5.7	0	0	0	7	5.2	0	0	0	0.248	96.3	116.0	0.09	1016
1	2019	1	1	0	10	0	5.8	0	0	0	7	5.1	0	0	0	0.247	95.54	117.87	0.09	1015
2	2019	1	1	0	20	0	5.9	0	0	0	7	5.1	0	0	0	0.246	94.88	119.75	0.09	1015
3	2019	1	1	0	30	0	6.0	0	0	0	7	5.1	0	0	0	0.246	94.23	121.64	0.09	1015
4	2019	1	1	0	40	0	6.0	0	0	0	7	5.7	0	0	0	0.245	97.77	123.53	0.09	10

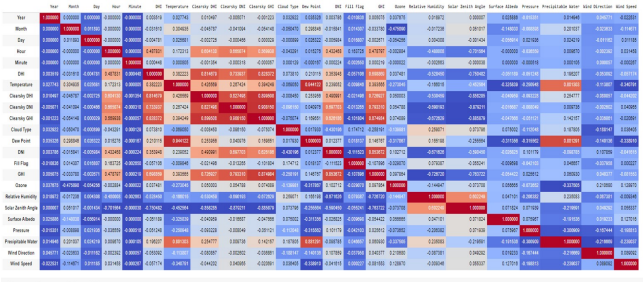
(Some sample data from NREL's data set)

Solar irradiance comprises three distinct types: Direct Normal Irradiance (DNI), which quantifies direct sunlight; Global Horizontal Irradiance (GHI), representing total sunlight on a horizontal surface; and Diffuse Horizontal Irradiance (DHI), indicating scattered sunlight reaching a horizontal surface. A basic EDA of the data shows high correlation with features like GHI , which is expected as the formula for calculating Global Horizontal Irradiance (GHI) is given by:

$$GHI = DNI \cdot \cos(Z) + DHI$$

Where:
GHI = Global Horizontal Irradiance

DNI = Direct Normal Irradiance
Z = Solar Zenith Angle
DHI = Diffuse Horizontal Irradiance

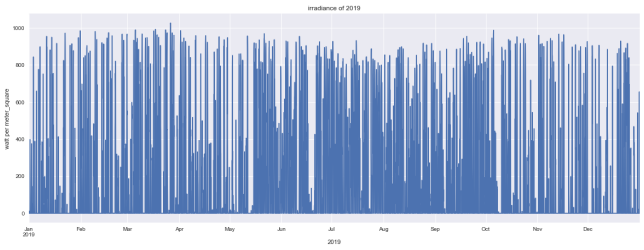


(Correlation Plot)

This time series analysis is uni-variate, concerned with DNI as it is vital for optimizing solar energy systems and forecasting energy generation, ensuring efficient utilization of renewable resources and grid integration.

5.2 Time Series Analysis

Utilizing a 10-minute interval DNI, the time data is used while discarding other columns. To assess the data's stationarity, the Augmented Dickey-Fuller (ADF) test [5] was used. The null hypothesis:
H0: given time series is non-stationary,
H1: given time series in stationary.
Employing the 'statsmodel'[6] library, the ADF test yielded an ADF statistic of -23.453798, a p-value of 0.0001, and critical values of 1% : -3.430, 5% : -2.862, and 10% : -2.567. Therefore, H0 is rejected, this implies the time series is indeed stationary.



(Irradiance in the year of 2019)

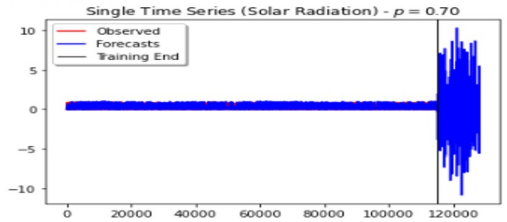
Now, the algorithm described above for this data is used, [7] Firstly, data is normalized using the Min-Max scaling technique to ensure all values lie within a common range. Some training data entries are randomly hidden to simulate missing data.A small fraction of these hidden entries are further consecutively hidden to create a more realistic scenario. Imputation is performed using two different techniques: Singular Value Decomposition (SVD) and Alternating Least Squares (ALS). Imputation involves filling in missing values with estimated values to create a more complete dataset. The Root Mean Squared Error (RMSE) values reported after imputation compare the imputed values with the original observed values in the training dataset.
The RMSE value of approximately 0.178151 indicates the accuracy of the imputation performed using SVD. This value represents the average difference between the imputed values and the true observed values in the training data. A lower RMSE indicates better agreement between imputed and observed values.
The RMSE value of around 0.101272 after ALS imputation indicates even better performance in terms of imputing missing values. The ALS algorithm iteratively refines the imputed values to minimize the error, leading to improved accuracy compared to SVD.
Now the next step is Forecasting:
After imputing missing values, the algorithm proceeds to generate forecasts for a larger number of points in the test dataset. The RMSE value of approximately 1.188100 represents the accuracy of the forecasts made by the forecasting model. This value shows the average difference between the predicted values and the actual observed values in the test data.

A higher RMSE in forecasting suggests that the model's predictions have a larger average error when compared to the true observed values. This could indicate areas where the forecasting model might be improved or where the data may exhibit more complexity.

The imputed data is compared with the true training data for evaluation using Root Mean Squared Error (RMSE). The observed and forecasted data are plotted against time, along with the

end of the training period marked. This visual representation helps understand the performance of the imputation and forecasting methods.

```
************************************************************
************************************************************
********** Running the Testing Scripts. **********
------------------- Test # 1 (Single TS). -------------------------
Training the model (imputing)...
SVD
 RMSE (training imputation vs obs)  = 0.178151
ALS
Iteration 1, Err = 0.0181, DeltaErr = inf
Iteration 11, Err = 0.0050, DeltaErr = 0.0000
Iteration 21, Err = 0.0050, DeltaErr = 0.0000
Iteration 31, Err = 0.0050, DeltaErr = 0.0000
Total Iterations = 34
63
 RMSE (training imputation vs obs)  = 0.101272
Forecasting (#points = 12800)...
 RMSE (prediction vs obs)  = 1.188100
Plotting...
```




```
********** Testing Scripts Done. **********
************************************************************
************************************************************
```

(output generated by the algorithm)

6 Conclusion

In conclusion, this projects investigated an algorithmic breakthrough in time series analysis. Leveraging the power of matrix estimation techniques and a novel approach to regression, the algorithm offers a versatile, adaptable, and robust solution for imputation, prediction, and dynamic state recovery. The application of this algorithm to solar irradiance prediction serves to highlight its efficacy in real-world contexts. By seamlessly merging theoretical advancements with practical utility, this algorithm advances the field of time series analysis, paving the way for impactful applications across diverse domains.

7 Future Work

In future work, it would be interesting to explore the application of the matrix estimation algorithm in multivariate time series analysis. This approach holds the potential to significantly enhance the accuracy and reliability of predictions, particularly in the context of solar irradiance forecasting.

Utilizing multivariate analysis with the matrix estimation algorithm offers a way to leverage the relationships and dependencies among various variables. By carefully investigating the correlations and interactions between different features, the algorithm's ability to capture intricate patterns within the data can be improved.

To achieve success in multivariate analysis, a thorough understanding of the correlations between different features is essential. This investigation will not only involve statistical analysis but also data visualization techniques. Moreover, deepening the understanding of the solar irradiance domain itself, including factors such as meteorological conditions, geographical location, and time of day, will guide the selection of relevant features and the formulation of meaningful models.

One specific application of this approach could involve short-term load forecasting of solar irradiance [8]. By incorporating correlated variables and their impact on solar irradiance, a more comprehensive prediction model can be developed. The integration of the matrix estimation algorithm into this multivariate context has the potential to provide accurate short-term forecasts that are both data-driven and informed by domain-specific insights.

Ultimately, the integration of multivariate analysis with the matrix estimation algorithm represents a promising direction for enhancing the predictive capabilities of models. By exploring correlations, deepening domain knowledge, and applying this enhanced methodology to short-term solar irradiance forecasting, more accurate and reliable predictions can be achieved, ultimately advancing the field of time series analysis.

References

[1] Anish Agarwal and Muhammad Jehangir Amjad and Devavrat Shah and Dennis Shen Model Agnostic Time Series Analysis via Matrix Estimation *arXiv*, 2019. 2, 3, 4

[2] A.A.H Damen, P.M.J Van den Hof, and A.K Hajdasinskit. Approximate realization based upon an alternative to the Hankel matrix: the Page matrix. Systems and Control Letters *Systems and Control Letters*, 1982. 2

[3] Nina Golyandina and Anton Korobeynikov and Alex Shlemov and Konstantin Usevich Multivariate and 2D Extensions of Singular Spectrum Analysis with thebRssa/bPackage *Journal of Statistical Software*, 2015. 2

[4] Sengupta, M., Y. Xie, A. Lopez, A. Habte, G. Maclaurin, and J. Shelby The National Solar Radiation Data Base (NSRDB) *Renewable and Sustainable Energy Reviews*,2018. 4

[5] Dickey, D. and W. Fuller Distribution of the Estimators for Autoregressive Time Series with a Unit Root *Journal of the American Statistical Association*, 1979. 5

[6] Seabold, Skipper, and Josef Perktold. tslib - A Time Series Library *https://github.com/jehangiramjad/tslib* 5

[7] Jehangir Amjad Statsmodels: Econometric and statistical modeling with python. *Proceedings of the 9th Python in Science Conference*,2010 5

[8] Akhtar S, Shahzad S, Zaheer A, Ullah HS, Kilic H, Gono R, Jasiński M, Leonowicz Z. Short-Term Load Forecasting Models: A Review of Challenges, Progress, and the Road Ahead. *Energies*,2023. 6