Time Series Forecasting: Basics

Jaemin Yoo

School of Electrical Engineering

Kim Jaechul Graduate School of AI

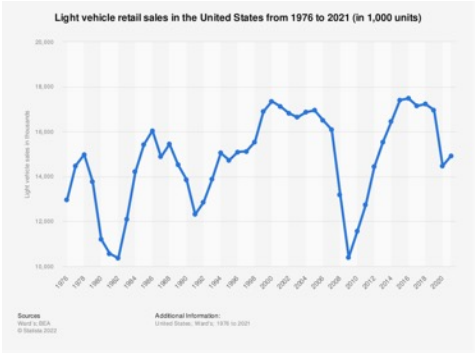
Outline

1. **<u>Introduction</u>**
2. Modeling choices
3. Linear regression
4. Summary

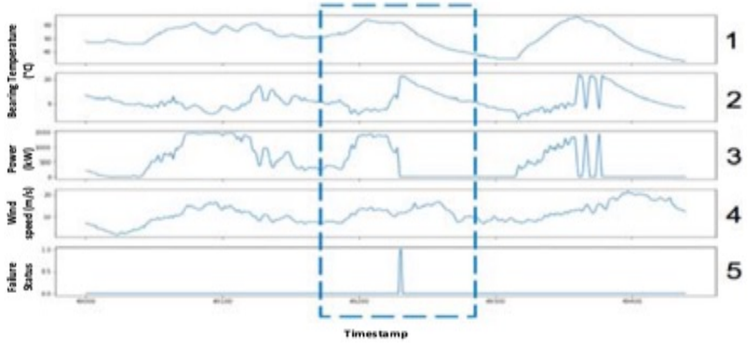
Time Series are Everywhere

- Any sequential data is **time series** whether it is ⋯
  - Fixed or variable length
  - With or without explicit timestamps
  - Univariate or multivariate data
  - Regular or irregular observations

Stock prices



Sales



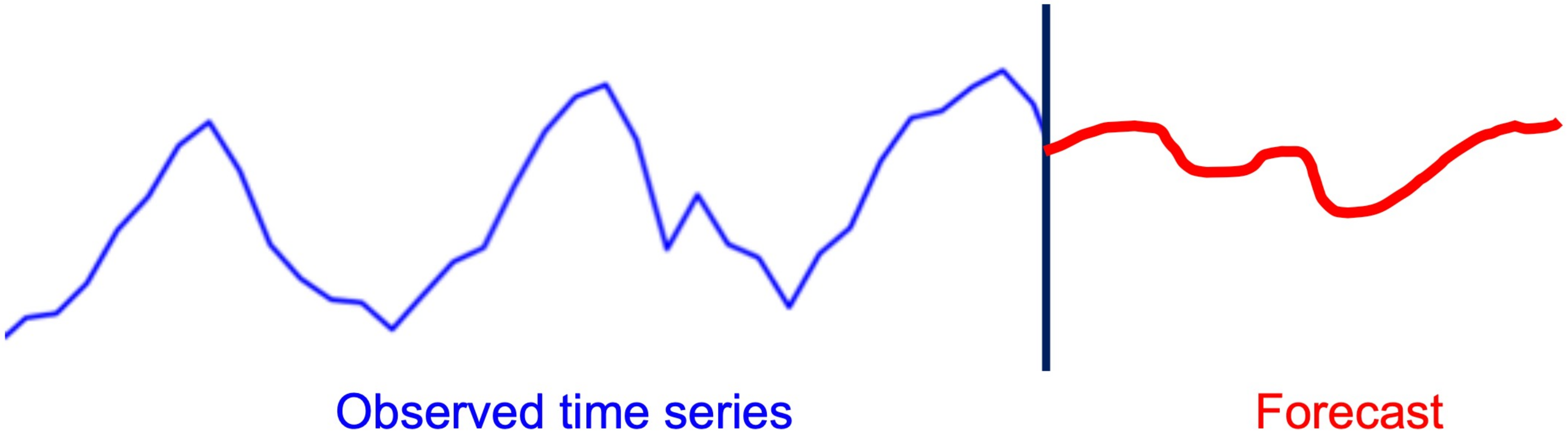
Sensors

Time Series Analysis

- **Time series analysis** is to solve problems defined on time series.
- **Time series-level problems:**
    1. Time series classification (ECG data → healthy or not)
    2. Time series anomaly detection (ECG data → something wrong)
    3. Time series clustering (ECG data → patient groups)
- **Observation-level problems:**
    1. Time series forecasting (stock prices → future prices)
    2. Time series forecasting as classification (stock prices → up/down)
    3. Abnormal event detection (stock prices → suspicious trades)

Time Series Forecasting

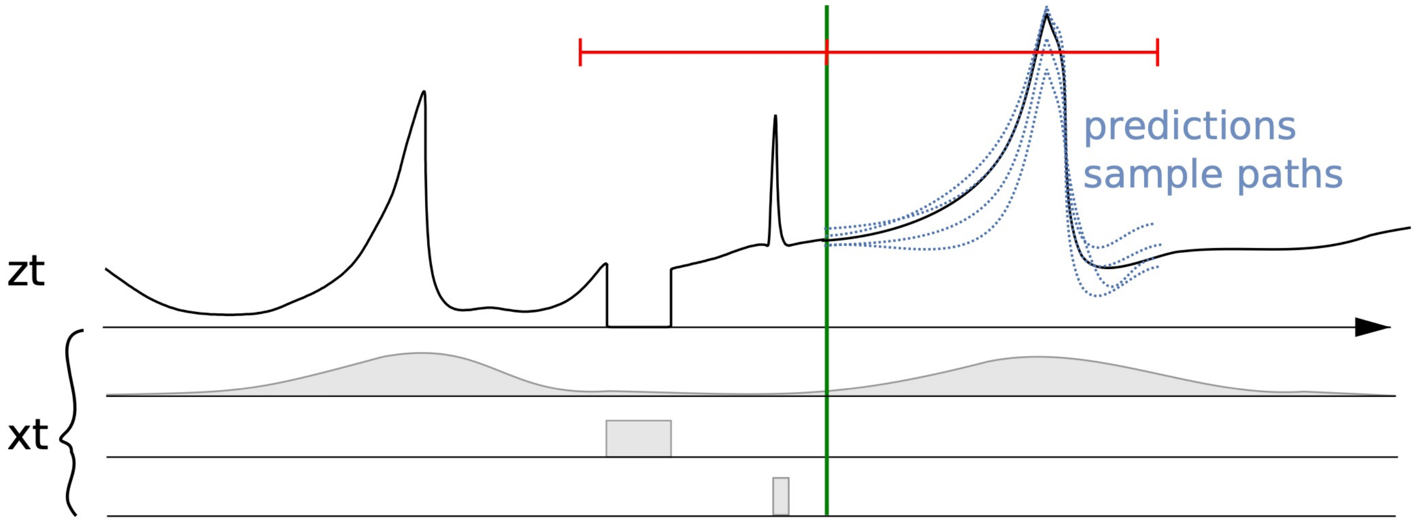
- We will study **time series forecasting** in this lecture.
  - A popular problem which is related to many practical applications.
  - Requires a deep understanding on the nature of time series.
  - Good forecasting models can be used for other problems as well.



Forecasting Problems: General Setup

- Let $i \in I$ be an item, and T be the current timestamp.
- **Setup:** Predict the future behavior of a time series $z_{i,t}$ given its past:

$$z_{i,0}, z_{i,1}, \cdots, z_{i,T} \Longrightarrow P(z_{i,T+1}, z_{i,T+2}, \cdots, z_{i,T+h}).$$



Forecasting Problems: General Setup

- **Point 1:** Predicting the distribution.
  - Our goal is to estimate the **distribution** of future behavior:

$$P(z_{i,T+1}, z_{i,T+2}, \cdots, z_{i,T+h}).$$

  - Instead, we assume to make **point forecasts** for simplicity.

$$\hat{z}_{i,T+1}, \hat{z}_{i,T+2}, \cdots, \hat{z}_{i,T+h}.$$

    - **Underlying assumption:** $P(z_{i,t}) = \mathcal{N}(\hat{z}_{i,t}, \sigma^2)$ where σ is a constant.
    - That is, we assume a Gaussian distribution with fixed standard deviation.

Forecasting Problems: General Setup

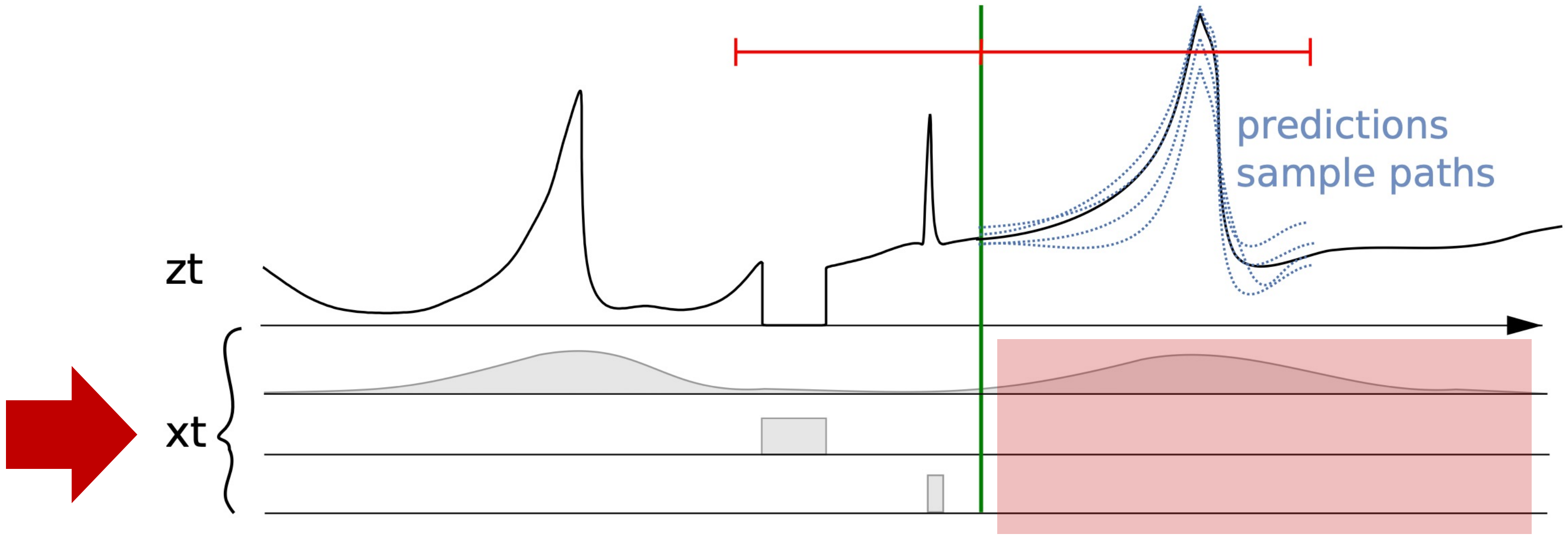
- **Point 2:** Predicting the sequence.
  - Our goal is to estimate the $\boldsymbol{h}$ **future steps** of future behavior:

$$\hat{z}_{i,T+1}, \hat{z}_{i,T+2}, \cdots, \hat{z}_{i,T+h}.$$

  - **Typical approach:** Predict the values in an **autoregressive** way.
    - Create a model f that predicts only one future step, i.e., $z_{i,T+1}$.
    - Apply f multiple times, e.g., use $\hat{z}_{i,T+1}$ to create $\hat{z}_{i,T+1}$, and so on.

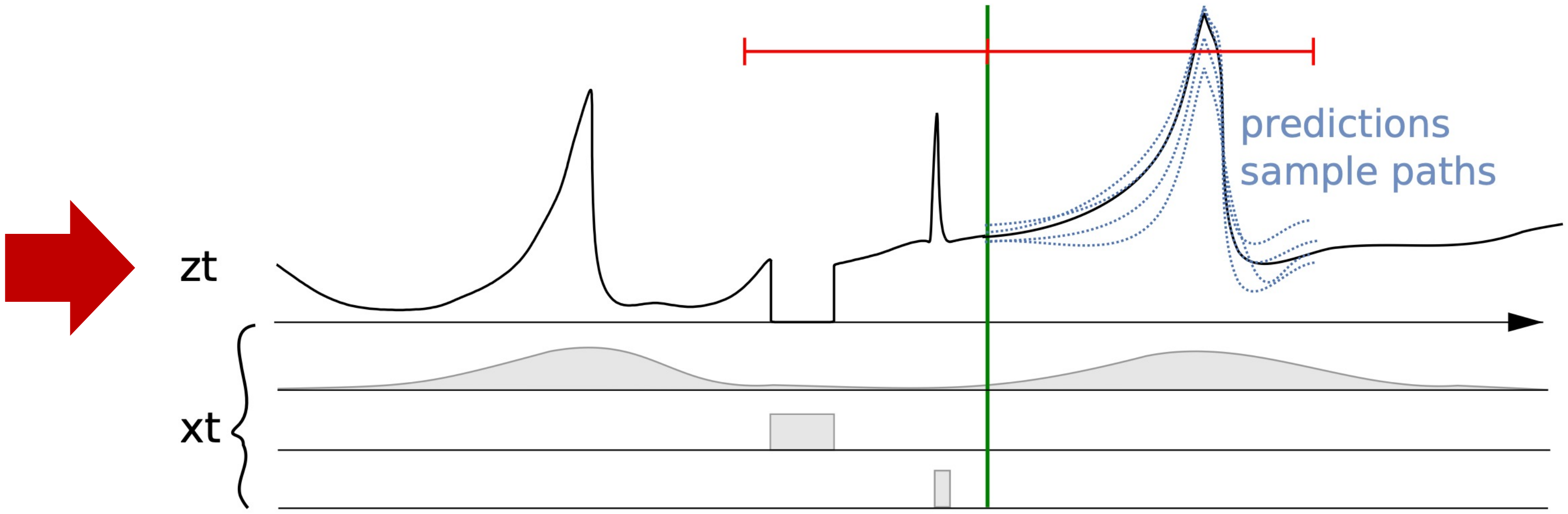
Forecasting Problems: General Setup

- **Point 3:** The existence of external attributes.
  - Better performance if an **attribute** $\boldsymbol{x}_{i,t}$ is given at time $t \in [1, T]$.
  - Autoregressive models require **future values** as well: $\boldsymbol{x}_{i,T+1}, \cdots, \boldsymbol{x}_{i,T+h}$.
    - If not, we need to use the *encoder-decoder* structure (later).



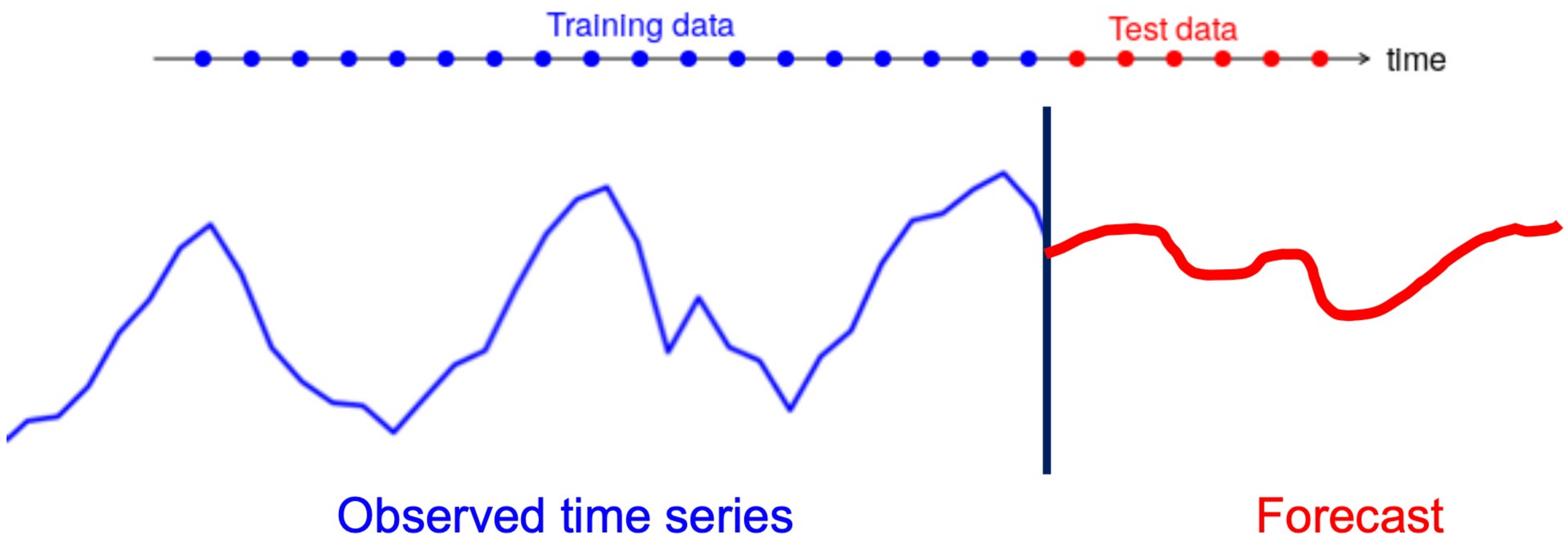
Forecasting Problems: General Setup

- **Point 4:** Univariate/multivariate time series.
  - We often want to predict multiple time series together.
  - **Multivariate models** are designed for the purpose.
    - E.g., predict the prices of Samsung Electronics and SK Hynix together.



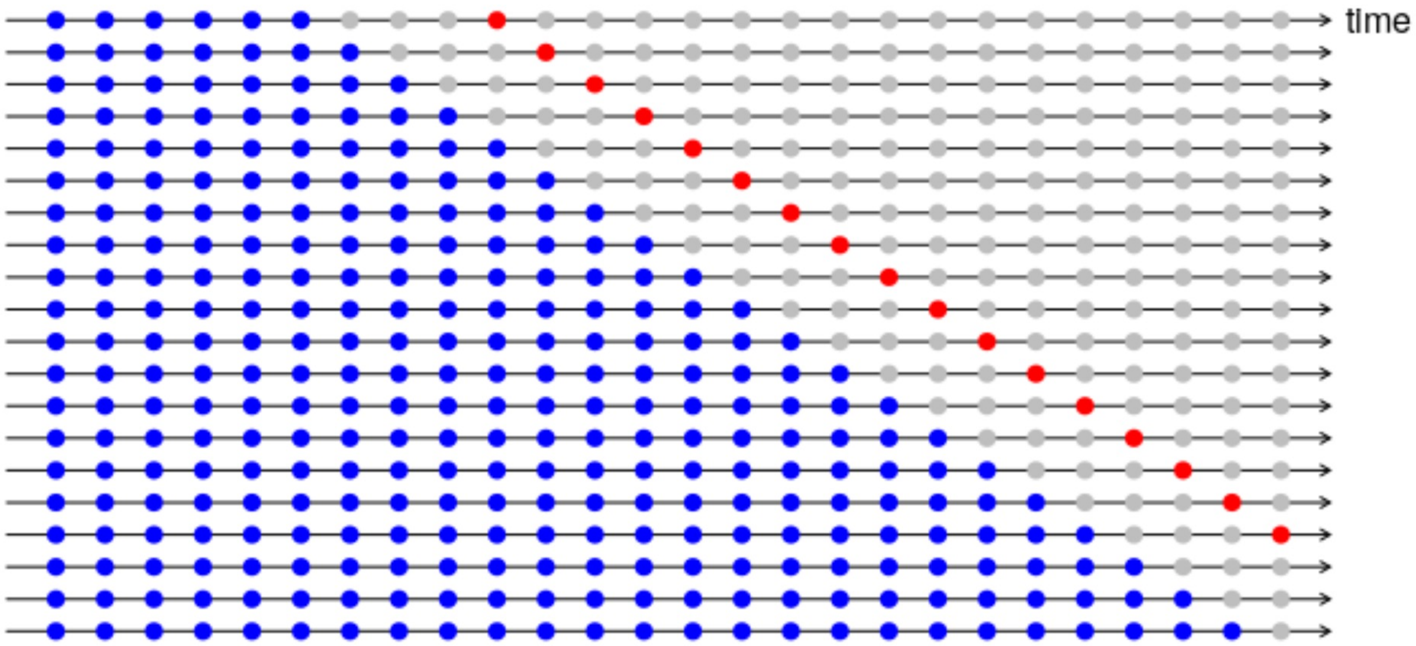
Training and Test Data

- We split an observed time series into **training** and **test data**.
- **Training:** We train a forecasting model f using training data.
- **Test:** We apply f to test data and evaluate its accuracy.



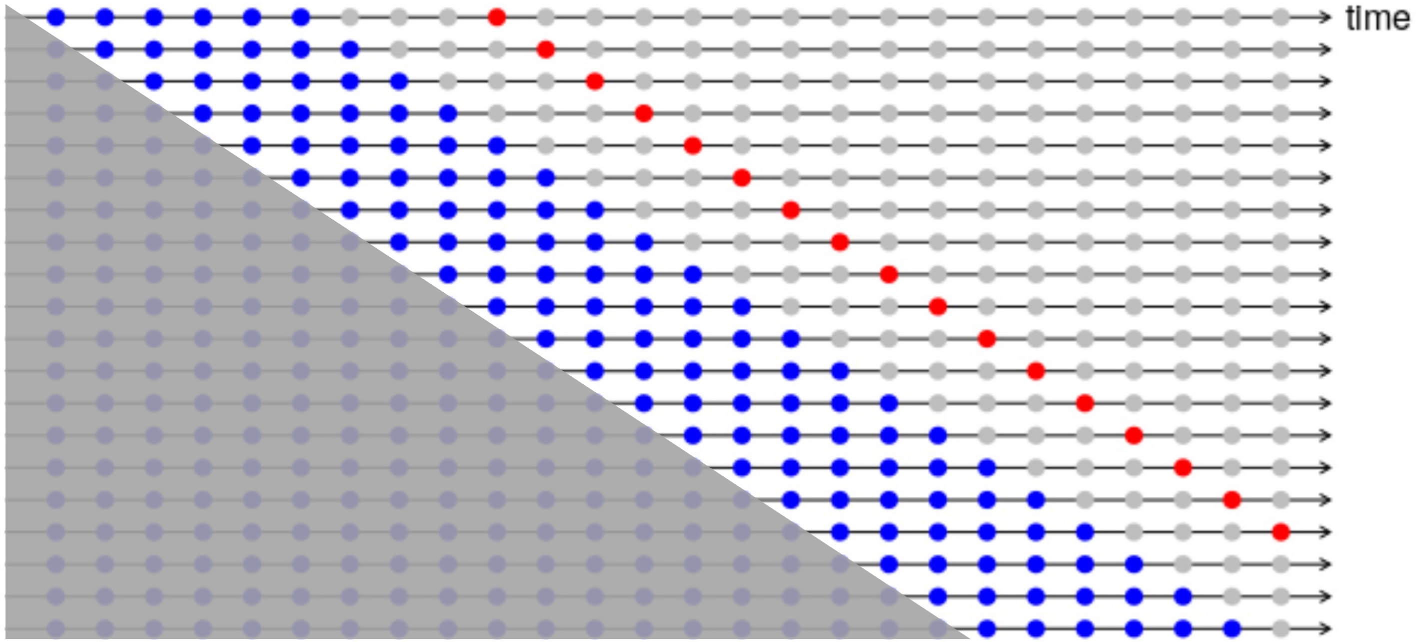
Training: Sliding Window

- After the split, we have a single (long) time series for training.
- We create **labeled training pairs** of short TS by **sliding window.**



Training: Sliding Window

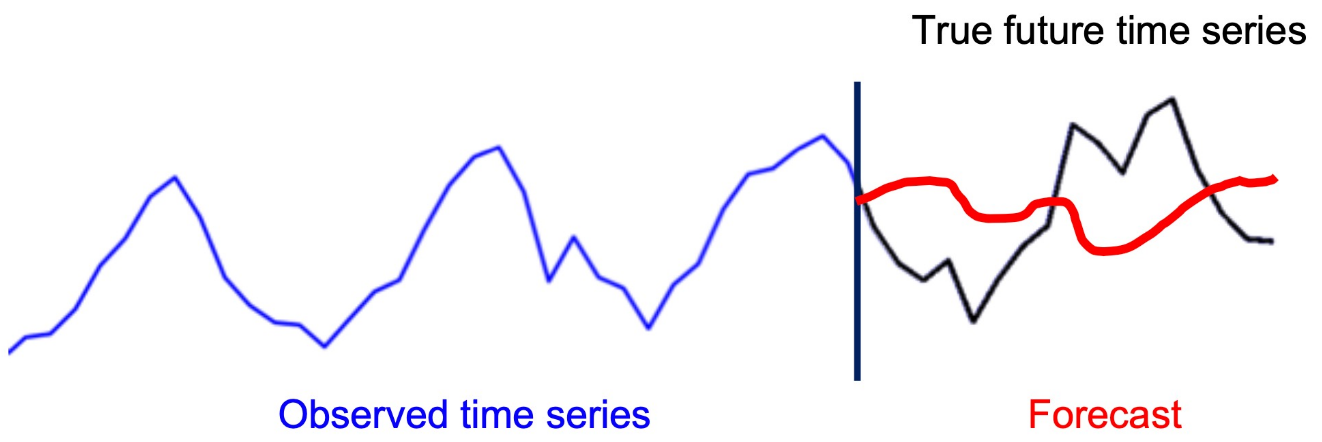
- In many cases, we fix the **window size** in all data (here $w = 6$).
- Here a prediction offset is 3, but it is not assumed in many cases.



Evaluation: Error Function

Note: i is skipped if obvious.

- After the training, an evaluation is done for test data.
  - Let $e_t = |z_t - \hat{z}_t|$ be the **absolute error** for each point z_t.
  - Mean absolute error (MAE): $1/h \cdot \sum_t e_t$.
  - Mean absolute percentage error (MAPE): $1/h \cdot \sum_t e_t/|z_t|$.
  - Root mean square error (RMSE): $\mathrm{sqrt}(1/h \cdot \sum_t e_t^2)$.



Evaluation: Remarks on Accuracy

- Potentially we can have three different accuracy measures:
  1. Loss function for training the model.
  2. Forecast accuracy metric for backtesting.
  3. Forecast accuracy measure for reporting to stakeholders.
- More accurate forecasts may not lead to better decisions.
- Need to carefully choose an evaluation metric for each step.

Outline

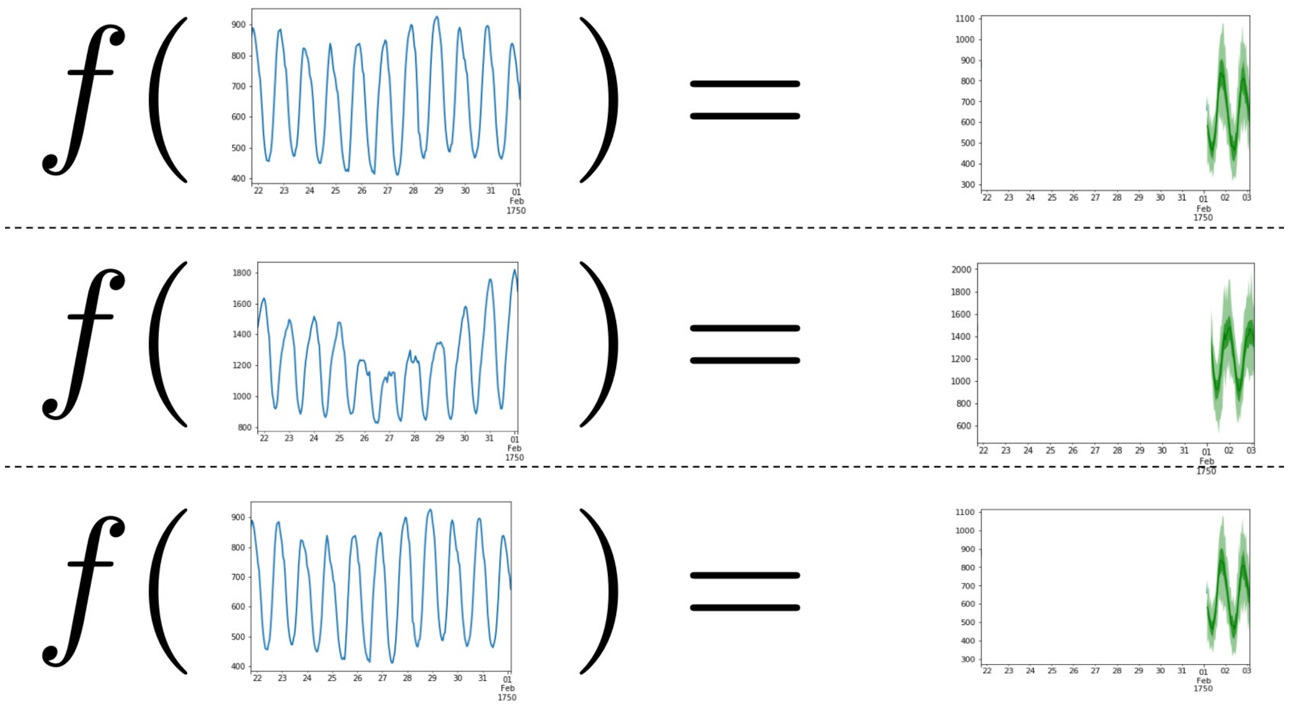
1. Introduction
2. **Modeling choices**
3. Linear regression
4. Summary

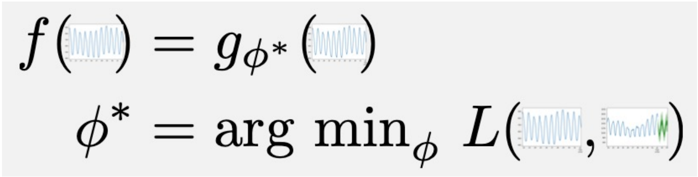
Modeling Choices

- Suppose that we have N time series of length L (ignoring $\boldsymbol{X}$).
- **Q1:** Do we need N different models or a single global model?
- **Q2:** Should we consider the relationships between N variables?

Local Univariate Model

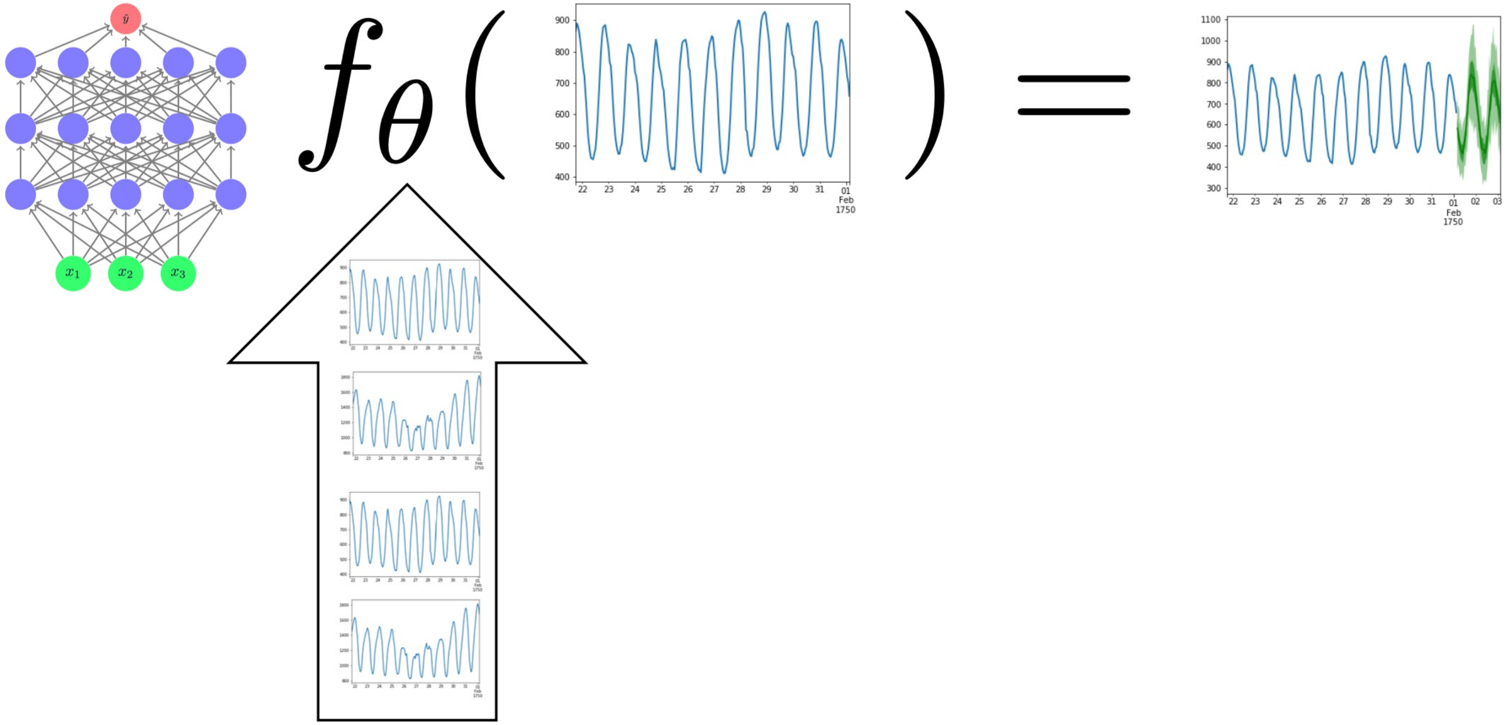
- **Local univariate model** predicts each TS instantly and separately.
  - Almost no training step is needed; the parameters are easily found.





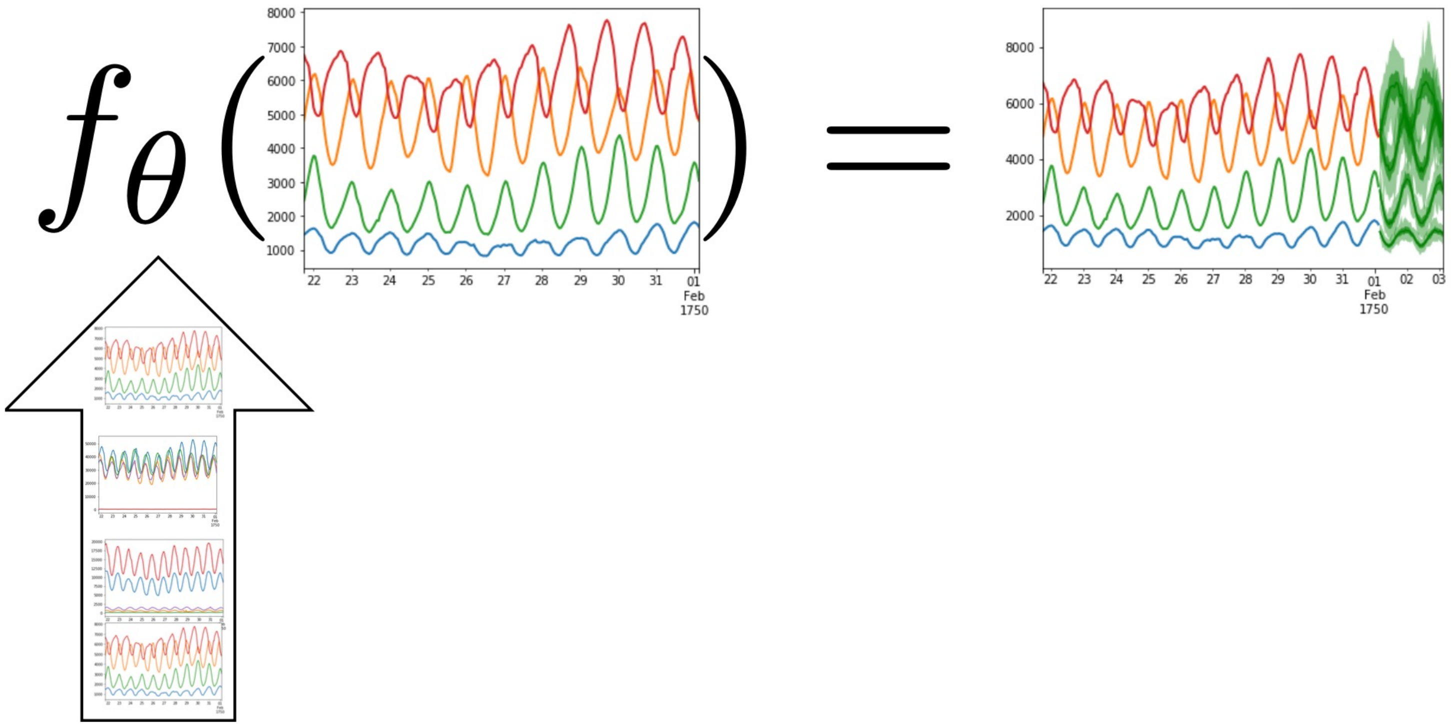
Global Univariate Model

- **Global univariate model** is trained once for all TS variables.
  - The trained model then works for each time series.



Multivariate Model

- **Multivariate model** takes/predicts all TS at the same time.
  - It considers the *relationships* between time series variables.



Training Pairs: Local Univariate Model

- We aim to create N different models.
- Each model uses only one of the N time series variables.
- Thus, we create $\mathcal{D}_i$ for the i-th model as follows:

$$\mathcal{D}_i = \{(z_{i,T-w+1}, \cdots, z_{i,T}, z_{i,T+1}, \cdots, z_{i,T+h}) | T \in [w, L-h]\}$$

= Input = Answer

- The size of training data is $|\mathcal{D}_i| = L - h - w + 1$.

Training Pairs: Global Univariate Model

- We aim to create one global model.
- The model uses any of the N time series variables.
- Thus, we create $\mathcal{D}$ as follows:

$$\mathcal{D} = \{(\underbrace{z_{i,T-w+1}, \cdots, z_{i,T}}_{= \text{Input}}, \underbrace{z_{i,T+1}, \cdots, z_{i,T+h}}_{= \text{Answer}}) | i \in [1, L] \text{ and } T \in [w, L-h]\}$$

- The size of training data is $|\mathcal{D}| = N(L - h - w + 1)$.

Training Pairs: Multivariate Model

- We aim to create one global model.
- The model uses all N time series variables at once.
- Thus, we create $\mathcal{D}$ as follows:

$$\mathcal{D} = \{(\mathbf{z}_{T-w+1}, \cdots, \mathbf{z}_T, \mathbf{z}_{T+1}, \cdots, \mathbf{z}_{T+h}) | T \in [w, L-h]\}$$

= Input = Answer

- The size of training data is $|\mathcal{D}| = L - h - w + 1$.

Remarks

- Global models are better than local models in many cases.
  - Both in terms of accuracy and stability.
  - Can learn knowledge shared across different time series.
- Multivariate forecasting models are not necessarily better.
  - The model becomes larger and more complex.
  - The number of training data decreases N times.

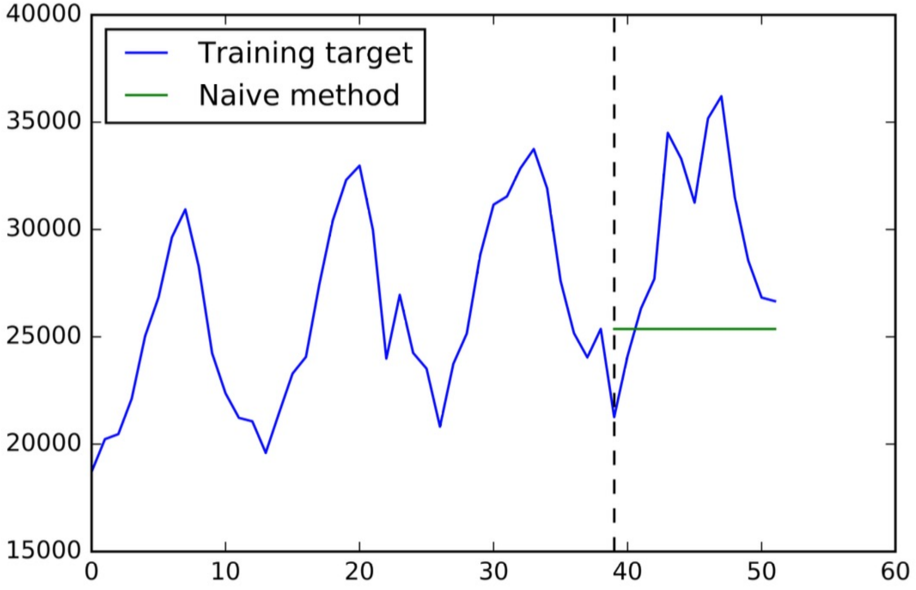
Outline

1. Introduction
2. Modeling choices
3. **Linear regression**
4. Summary

Parameter-Free Forecasting Models

- **Naive method:** Forecasts are equal to the last observed value:

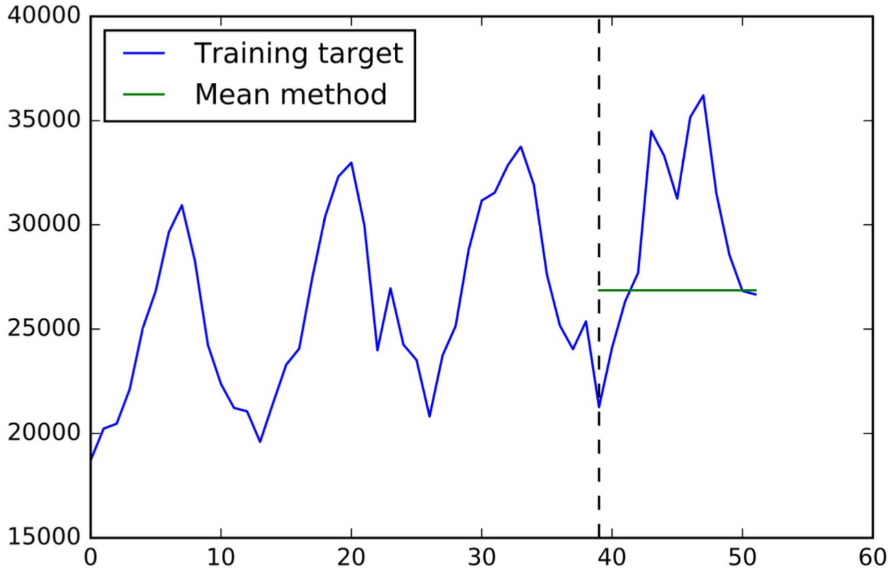
$$z_{T+t} = z_T,\ \forall t = 1, 2, \cdots, h.$$



Simple Forecasting Models

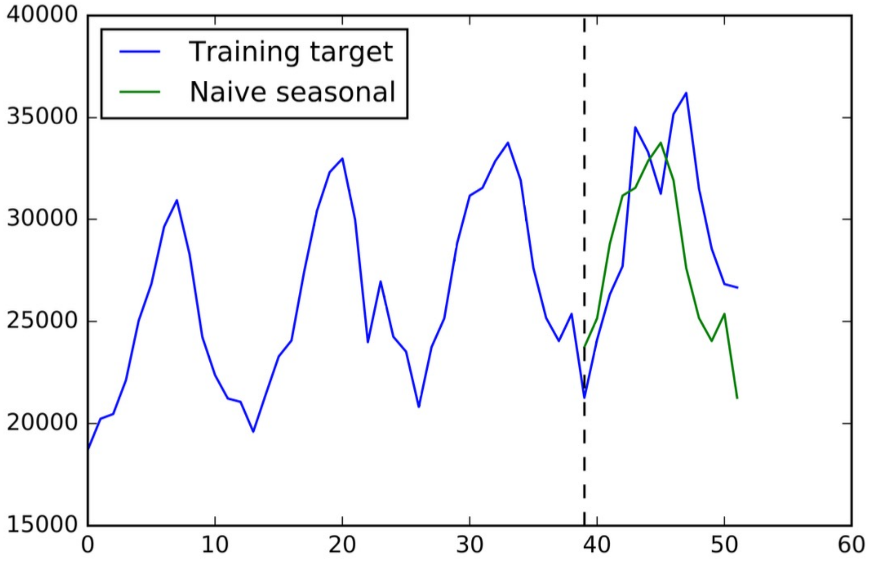
- **Mean method:** Forecasts are equal to the average of all observations:

$$z_{T+t} = \frac{1}{W}(z_{T-W+1} + z_2 + \cdots + z_T),\ \forall t = 1, 2, \cdots, h.$$



Simple Forecasting Models

- **Naive seasonal method:** Forecasts are taken from the last *season*.
  - How to capture the exact seasonality is another problem.
  - E.g., the same month of the previous year.

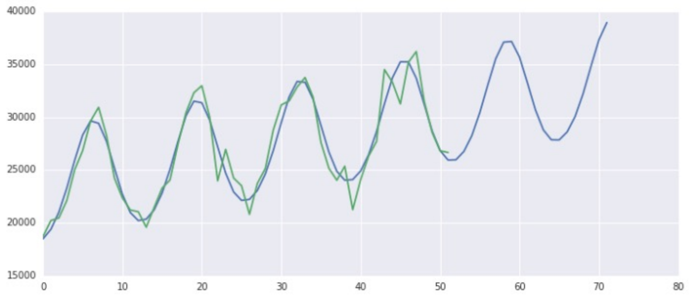


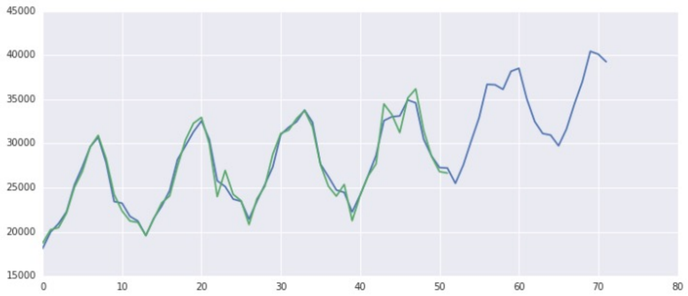
Forecasting with Linear Regression

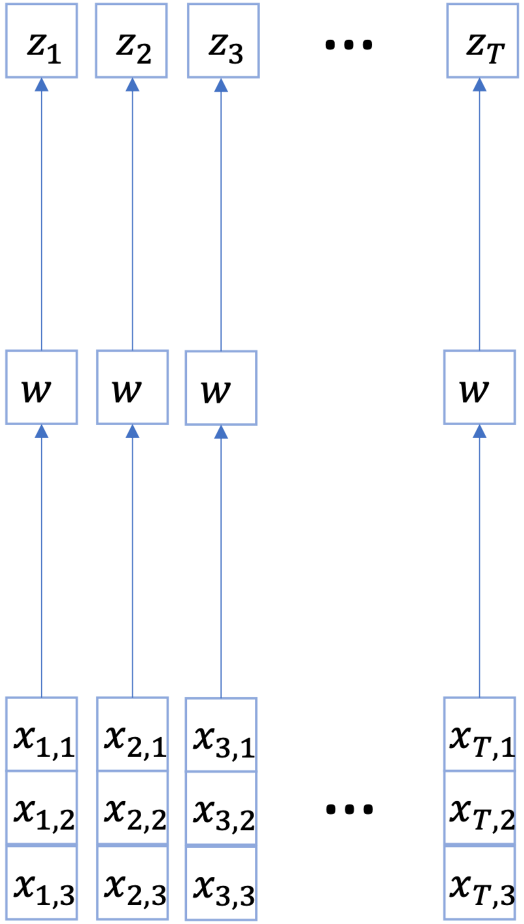
- **Linear regression:** Assume that a prediction $\hat{z}_t$ is a weighted combination of features $x_{t,1}, \cdots, x_{t,D}$:

$$\hat{z}_t = \sum_{d=1}^{D} w_d x_{t,d}.$$

  - Then, estimate the weights w_d through *training*.
  - The features $x_{t,d}$ can be defined in various ways.
    - Previous observations, additional information, etc.



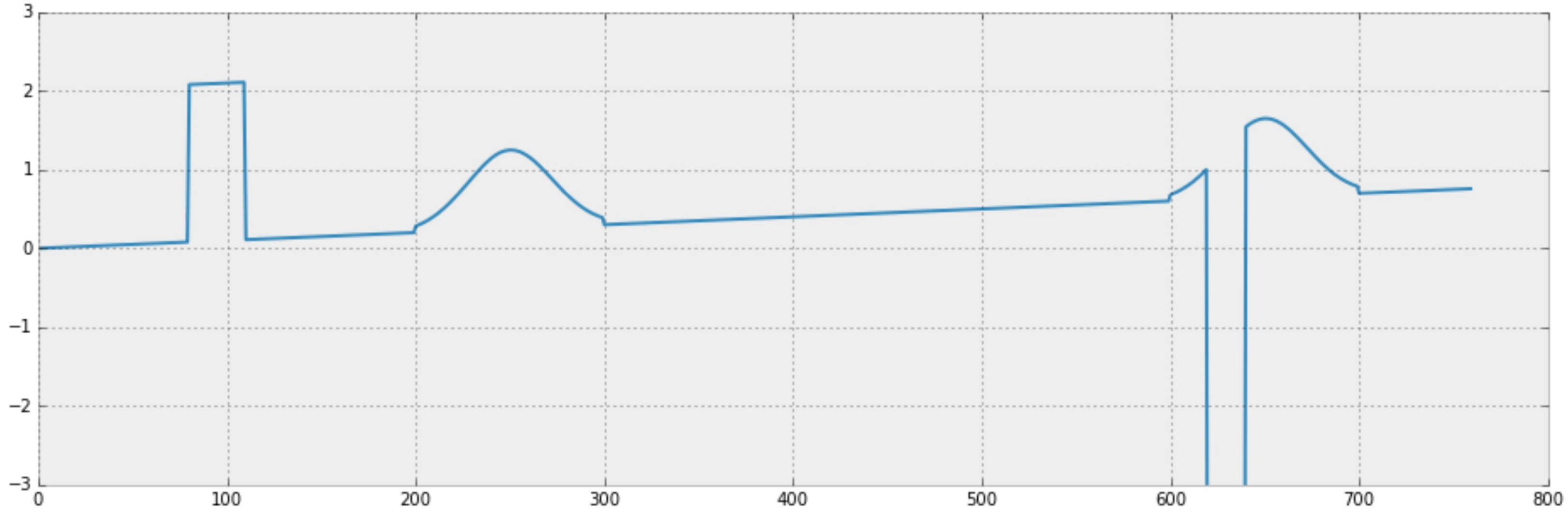




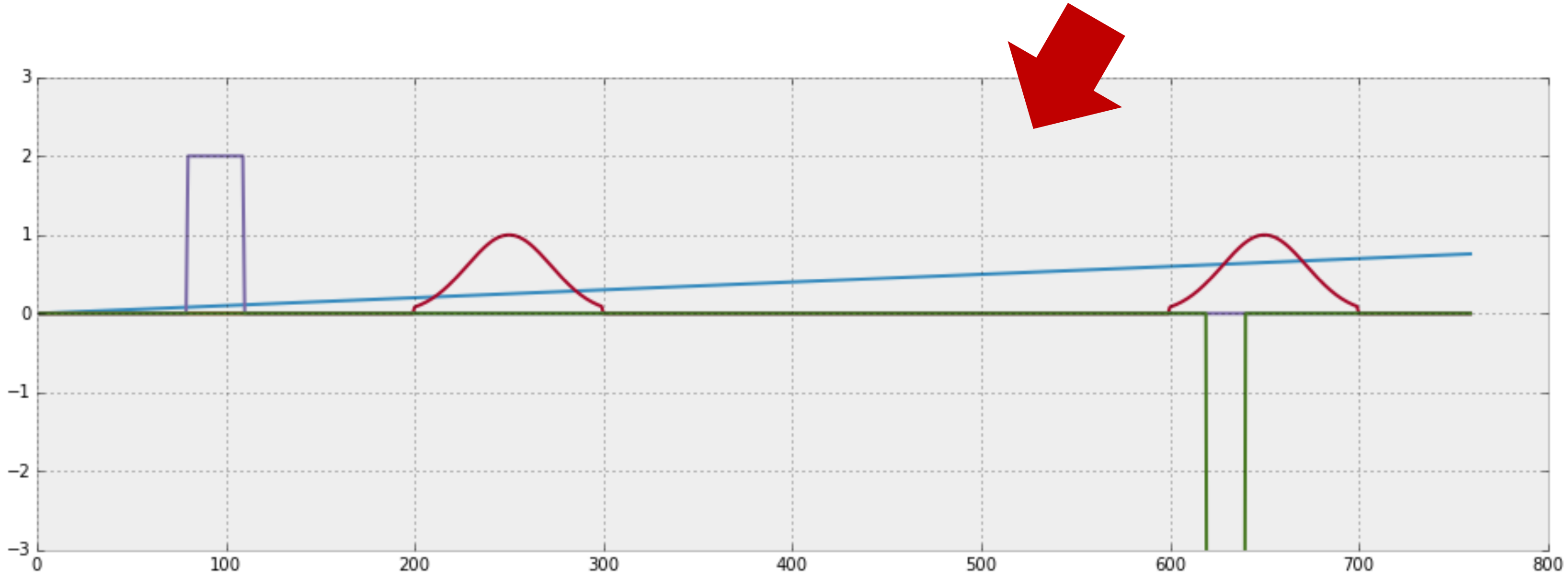
Features for Linear Regression

- The features for linear regression are themselves time series.
  - Since they are observed over time: $x_{1,d}, x_{2,d}, \cdots, x_{T+h,d}$.
- Possible features include the following:
  1. External attributes
  2. Lagged target values (e.g., z_{t-1} and z_{t-2} as features to predict z_t)
  3. Trend features (e.g., $z_{t-1} - z_{t-2}$ as a feature to predict z_t)
  4. Seasonal lagged target values (e.g., z_{t-S} as a feature to predict z_t)
  5. (Weighted) average features (e.g., $\mathrm{mean}(z_{t-7:t-1})$)

Examples



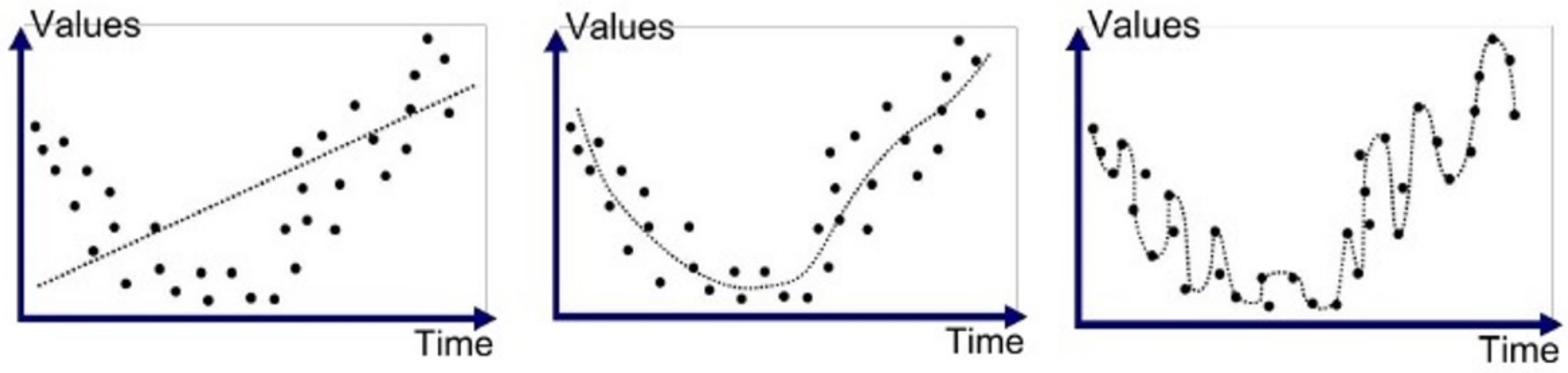
3
2
1
0
−1
−2
−3
0
100
200
300
400
500
600
700
800



3
2
1
0
−1
−2
−3
0
100
200
300
400
500
600
700
800

How to Choose Features

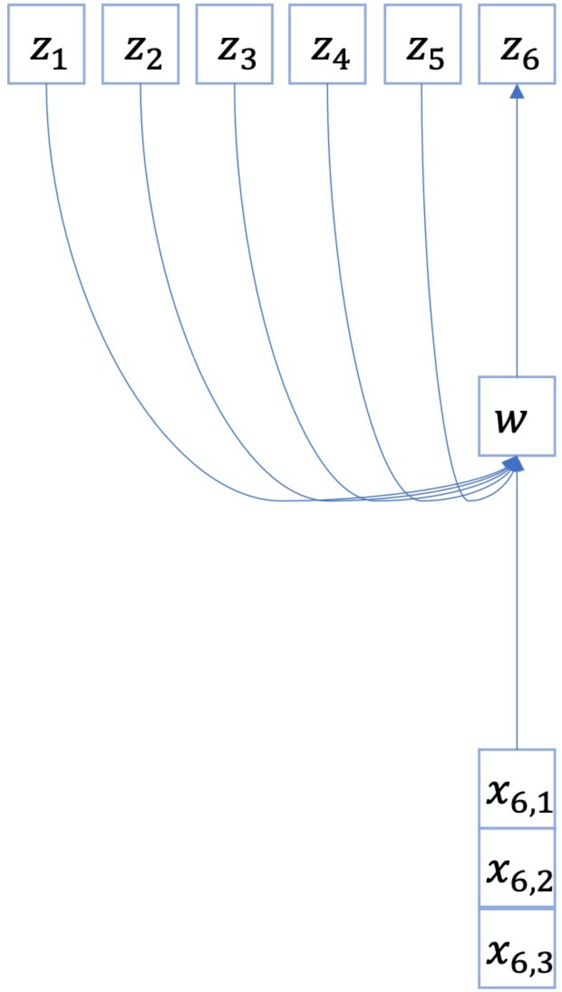
- **Q:** What if we include *all features* into linear regression?
- This is a classical example of **overfitting:**
  - Model starts fitting noise with too many free parameters.
  - Model is not generalizing well to unseen test data.



Autoregressive Models

- **Autoregressive (AR) models** focus on lagged values.
  - Use lagged values z_{t-l} as features for predicting z_t.
  - Also include two new terms b and ϵ.
    - b is a constant which we train along with the weights $\mathbf{w}$.
    - $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is a random noise which we cannot control.
- AR is defined as follows:

$$\hat{z}_t = \sum_{l=1}^{p} w_l z_{t-l} + b + \epsilon.$$

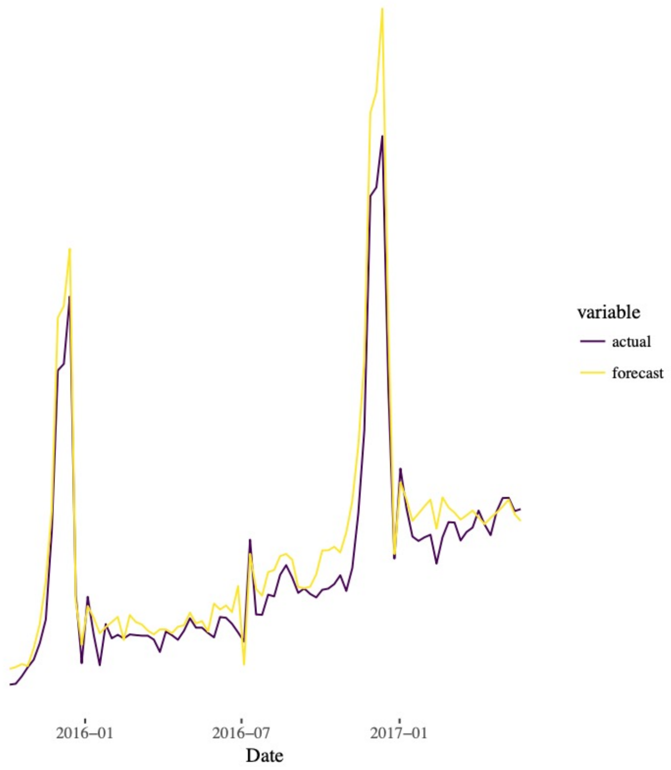


Outline

1. Introduction
2. Linear regression
3. State space models
4. **<u>Summary</u>**

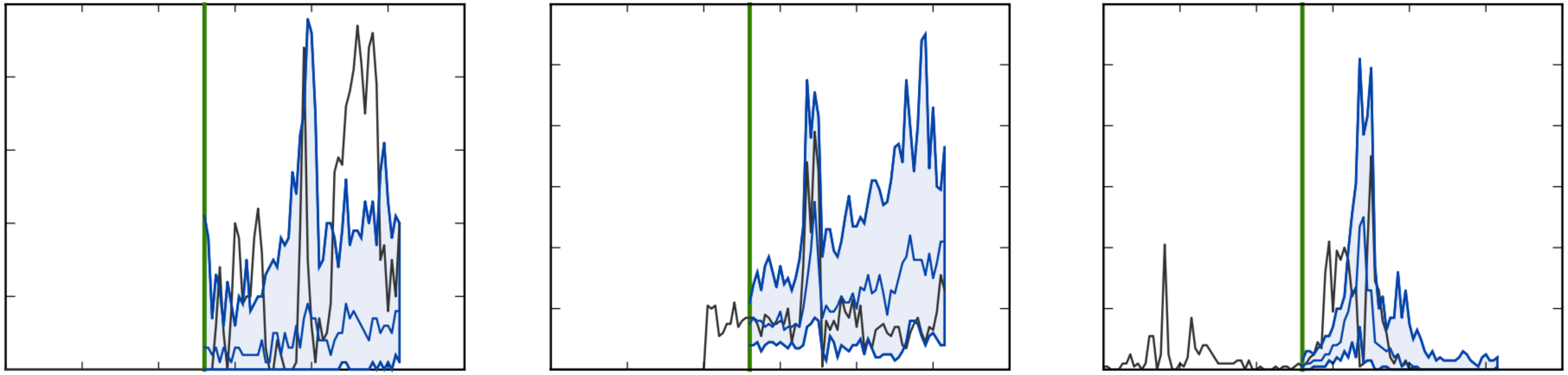
When to Use Classical Methods

- Classical methods are good for **strategic forecasting problems**.
- For example, to predict the overall Amazon retail demand years into the future.
- When time series have enough history, are regular and exhibit clear patterns.



When to Use Classical Methods

- Classical methods struggle with **operational forecasting problems**.
- For example, to predict the demand for each product.
- Time series are irregular and may not contain enough history.



Classical Methods: Pros and Cons

- **Pros:**
  - De-facto standard; widely used.
  - Decomposition → decoupling.
  - *White box:* explicitly model-based and thus interpretable.
  - Requires little resources to run.
- **Cons:**
  - Requires manual work by experts.
    → Hard to tune & maintain.
  - Cannot learn complex patterns.
  - *Model-based:* all effects need to be explicitly modeled.