Time series and Spatio-Temporal forecasting as a data imputation problem

Tuan Tran Xiao-Ping Zhang

Ryerson University

September 3, 2019

Overview

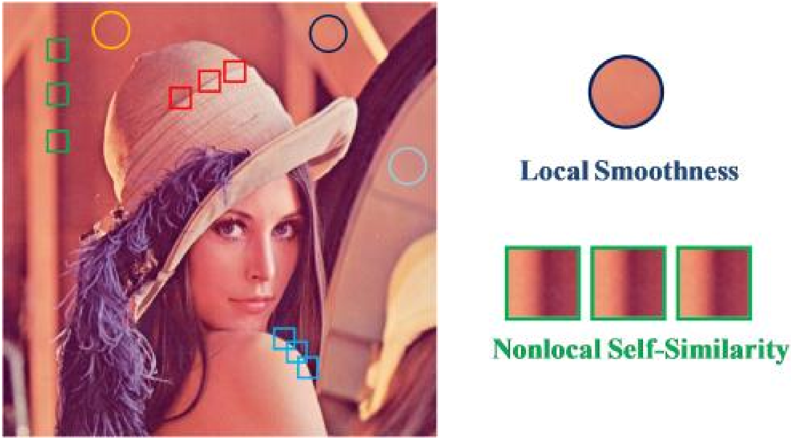
1. Introduction

2. Time series forecasting

3. Spatio-temporal forecasting

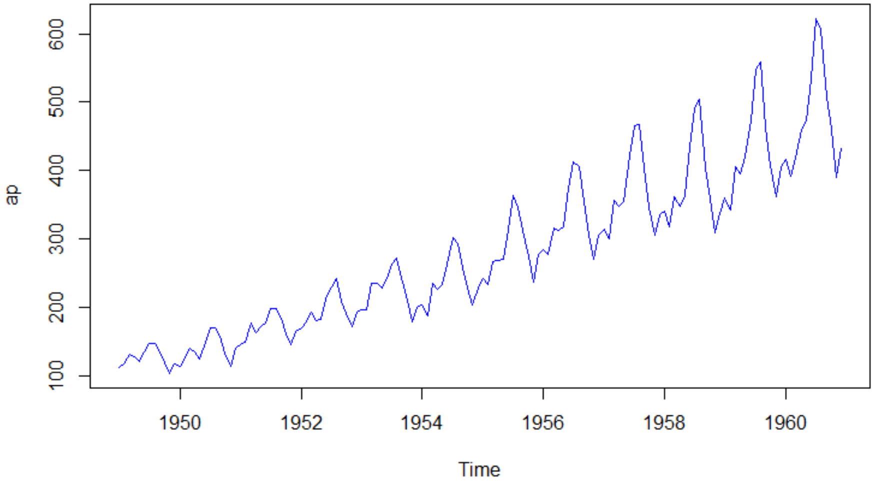
Non local self similarity (NLS)

NLS is a common approach for signal denoising (NLM, BM3D, low rank matrix etc.)



NLS in time series

Monthly Global Air Passenger Dataset (Box-Jenkins)



Non parametric regression

Problem: Given $(X_i, Y_i)_{i=1..n}$ and x, predict y.

- Average estimate: Take the mean of past outputs associated with **similar** inputs

$$y(x) = \frac{\sum_{i=1}^{n} I_{\|x-X_i\|\leq h} Y_i}{\sum_{i=1}^{n} I_{\|x-X_i\|\leq h}} \tag{1}$$

- Nadaraya-Watson estimate, assuming (X_i, Y_i) are drawn from the same distribution

$$y(x) = \frac{\sum_{i=1}^{n} K(\frac{x-X_i}{h}) Y_i}{\sum_{i=1}^{n} K(\frac{x-X_i}{h})} \tag{2}$$

  where K denotes the kernel function, for example truncated Gaussian kernel

$$K(x) = e^{-\frac{x^2}{2}} 1_{|x|\leq a}$$

Forecasting vs denoising

- Simple NW estimate is similar to NLM in signal denoising
  $$\widehat{X}_j = \sum_i W_{ij} X_i,$$
  where $\sum_i W_{ij} = 1$.
- Average or NW estimate is sensitive to outliers.
- We want to have a counterpart algorithm of BM3D for time series/spatio temporal forecasting.

Problem formulation

- We assume that the dynamics of the time series follows an autoregressive model

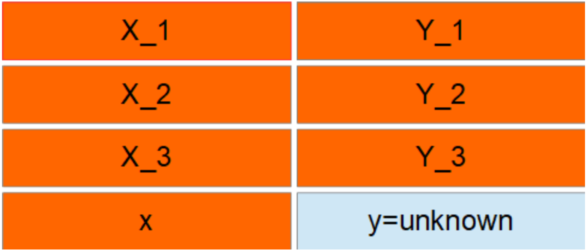
$$S_t = m(S_{t-1}, \dots, S_{t-L}) + \epsilon_t \tag{3}$$

- Given $(S_{t-s})_{s\geq 0}$ predicting a vector of future values $(S_{t+h})_{1\leq h\leq H}$
- Denote $x = (S_{t-l})_{0\leq l\leq L-1}$ and $y = (S_{t+h})_{1\leq h\leq H}$ (y is unknown). We denote (X_i, Y_i) the historical data.
- Similarity matching:

$$I_h(x) = \{s : d(X_s, x) \leq h\} \tag{4}$$

where d denotes a pseudo distance such as Pearson correlation or Euclidean distance.

Matrix completion



X_1	Y_1
X_2	Y_2
X_3	Y_3
x	y=unknown

Matrix completion without noise

- Denote A the matrix formed by stacking up the vectors $z = (x, y)'$ and $Z_i = (X_i, Y_i)'$ as columns:

$$A = A(y) = \begin{bmatrix} X_1' & \dots & X_n' & x' \\ Y_1' & \dots & Y_n' & y' \end{bmatrix} = [Z_1 \dots Z_n \, z] \tag{5}$$

- Matrix completion with exact observations (Ω)

$$\min_M \{\text{rank}(M) : \quad M_\Omega = A_\Omega\} = \min_y \{\text{rank}(A(y))\} \tag{6}$$

- The previous problem is NP-hard and non convex. A convex relaxation version is

$$\min_M \{\|M\|_* : \quad M_\Omega = A_\Omega\} \tag{7}$$

where $\|M\|_*$ denotes the nuclear norm.

Matrix completion is powerful!

Figure: 80% missing values

Figure: recovered image

Matrix completion (MC) with noise

- In practice, data are observed with noise. Moreover, when the missing ratio is small (which is the case of this work), the missing values do not have significant impact on the rank of the matrix.
- **Method 1. Low rank MC with noise**

$$\min_{M}\{\|M_{\Omega} - A_{\Omega}\|_F^2 : \quad \text{rank}(M) \leq r\} \tag{8}$$

- Notice that when there is no missing values, the solution is obtained by PCA: The rank r approximation of a matrix A is given by

$$A^r = U^r \Sigma^r (V^r)'$$

where $A = U \Sigma V'$ denotes the SVD of A.

Low rank MC with noise

Theorem

The predicted value $\hat{y}$ of problem 8 satifies the following fixed point equation

$$y = [A(y)]^r_{\Omega^c}$$

where Ω^c denotes the complement set of Ω.

- The above theorem gives an iterative algorithm to solve y: start with an initial guess for y (e.g. average estimate), then apply PCA on $A(y)$ to obtain an approximate matrix of rank r which give a new estimate of y. Repeat until convergence.
- This algorithm is monotonic. Indeed, if M_0 has rank r and corresponds to y_0, then this algorithm gives M_1 and y_1 such that

$$\|[M_1]_\Omega - A_\Omega\|_F^2 \leq \|[M_0]_\Omega - A_\Omega\|_F^2$$

Algorithm

Algorithm 1 Low rank matrix completion with noise

```
1: Initialize the missing values: $y = y_0$
2: Apply PCA on the matrix $A(y_0)$ to obtain the low rank matrix $M_1 = A(y_0)^r$ and the new estimate $y_1 = [M_1]_\Omega$
3: If $d := \|y_1 - y_0\| > \epsilon$ : replace $y_0$ by $y_1$ and repeat the previous step until $d \leq \epsilon$
4: Return $y_n$                                                            ▷
```

Hard and soft thresholding

- This algorithm does not ensure the uniqueness of solution nor the global optimality.
- If the rank is replaced by nuclear norm, then the PCA step is replaced by the soft thresholding operator.

$$M_1 = US^{\lambda}(\Sigma)V'$$

where S^{λ} replaces Σ_{ii} by $(\Sigma_{ii} - \lambda)^{+}$

Graph Laplacian Regularizer Approach

Method 2. Graph Laplacian Regularizer.

- Define the graph adjacency matrix $W = (W_{ij})$ where

$$W_{ij} = e^{-\gamma \|X_i - X_j\|^2}.$$

  And similarly $W_i = e^{-\gamma \|x - X_i\|^2}$.

- We ignore the inputs in the matrix A :

$$A(y) = [Y_1' \ldots Y_n' \, y'] = [Y' \, y'], Y \in \mathbf{R}^{n \times d}$$

  The objective function

$$\min_{\hat{Y}, y} \| \hat{Y} - Y \|^2 + \lambda \sum_{ij} W_{ij} \| \hat{Y}_i - \hat{Y}_j \|^2 + \lambda \sum_i W_i \| \hat{Y}_i - y \|^2$$

GLR approach

- Fixing $\hat{Y}$ we obtain the NW-like estimate

$$y = \frac{\sum_i W_i \hat{Y}_i}{\sum_i W_i} = \alpha \hat{Y}, \alpha \in \mathbf{R}^{1 \times n}$$

- Plug this into the objective function:

$$\min_{\hat{Y}, y} \|\hat{Y} - Y\|^2 + \lambda \sum_{ij} [W_{ij} + \frac{W_i W_j}{\sum_i W_i}] \|\hat{Y}_i - \hat{Y}_j\|^2$$

Explicit solution w.r.t the **modified Graph Laplacian**:

$$\hat{Y} = (1 + \lambda \tilde{L})^{-1} Y.$$

So y is weighted average of denoised outputs.

GLR approach

Lemma

The predicted value y is a weighted average of historical outputs Y, i.e.

$$y = \beta Y, \sum_{i=1}^{n} \beta_i = 1.$$

Notice that $\beta = \alpha(1 + \lambda\tilde{L})^{-1}$. Moreover, $\tilde{L}\mathbf{1} = 0$. Hence $\beta\mathbf{1} = \alpha\mathbf{1} = 1$.

Empirical results

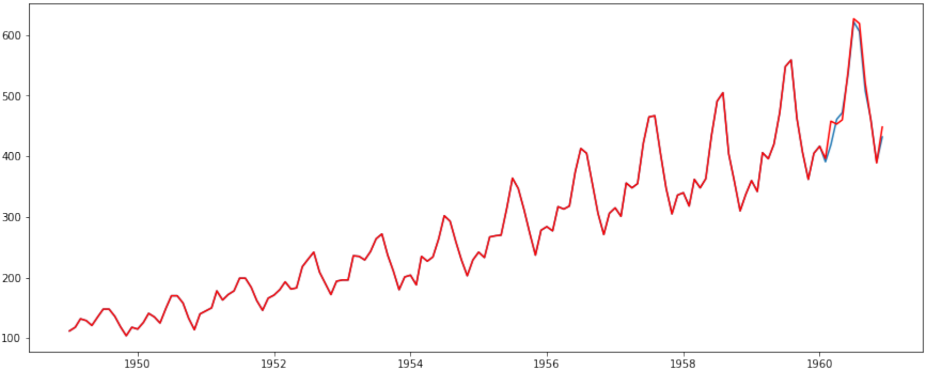
For illustration purposes, consider the set of Monthly air passengers, from 1949 to 1961. We test the predictive power of different models: our model and

- Parametric models: ARIMA
- Nadaraya-Watson with uniform kernel.
- Nuclear norm model
- Graph signal model.

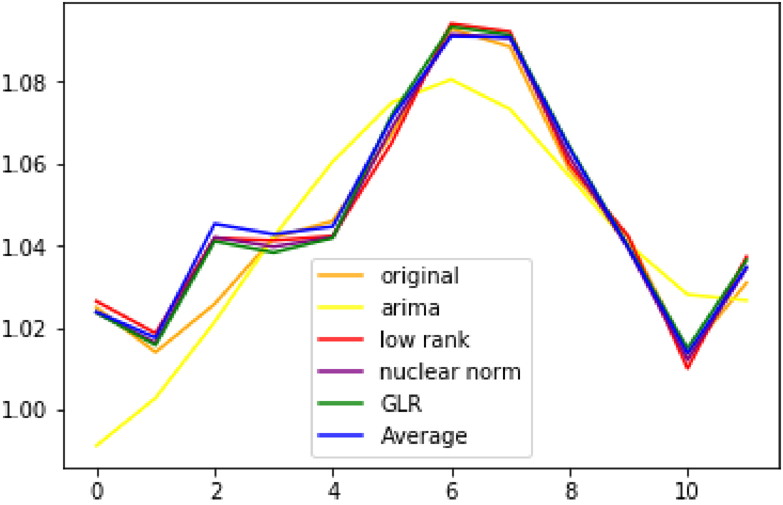
We use the relative root mean squared error to measure the predictive power of a given method

$$\hat{e}_T = \frac{100H}{\sum_{h=1}^{H} Y_{T+h}} \sqrt{\frac{\sum_{h=1}^{H}(\hat{Y}_{T+h} - Y_{T+h})^2}{H}}$$

Empirical results



Empirical results



Empirical results

Table: Relative Errors for different methods

Nuclear Norm	**PCA**	**GLR**	**Avg**	**ARIMA**
0.50%	0.53%	0.53%	0.60%	1.29%

Spatio-temporal forecasting: a technical extension

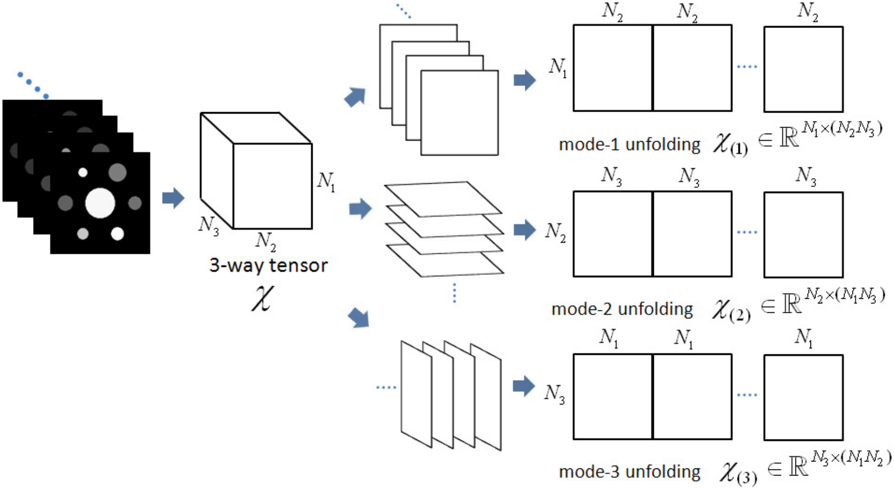
- Consider a spatio-temporal process of $d-$dimension

$$S_t = m(S_{t-1}, \ldots, S_{t-L}) + \epsilon_t \tag{9}$$

  Our goal is to predict S_{t+1}.

- By similar method as in the time series case, we come to a low rank tensor completion problem

Tensor



3-way tensor

$\mathcal{X}$

mode-1 unfolding $\mathcal{X}_{(1)} \in \mathbb{R}^{N_1\times(N_2N_3)}$

mode-2 unfolding $\mathcal{X}_{(2)} \in \mathbb{R}^{N_2\times(N_1N_3)}$

mode-3 unfolding $\mathcal{X}_{(3)} \in \mathbb{R}^{N_3\times(N_1N_2)}$

Tensor completion

Definition (Tensor completion with noisy observations)

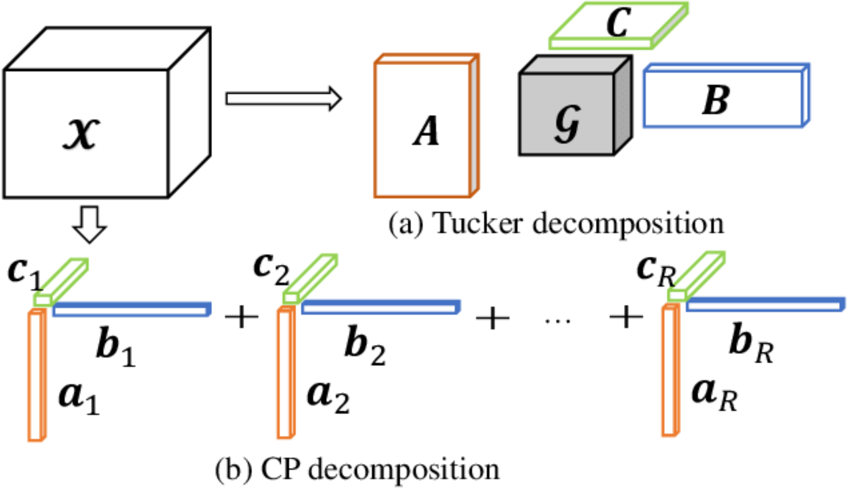
Given an incomplete tensor $\mathcal{T}$ with observed values on the set Ω, estimate the low rank tensor $\mathcal{X}$ such that

$$\min_{\mathcal{X}} \|\mathcal{X}_\Omega - \mathcal{T}_\Omega\|^2 \tag{10}$$

$$\text{subject to} \quad f(\text{rank}(\mathcal{X})) \leq r. \tag{11}$$

where the function f can be a vector-valued or scalar function.

Tensor decomposition



(a) Tucker decomposition

(b) CP decomposition

Tensor completion

- Low Tucker rank tensor completion

$$\|\Omega \circ (\mathcal{T} - \mathcal{G} \times_1 A_1 \times_2 A_2 \times_1 A_3)\|_F^2$$

  where $\mathcal{G}$ denotes the core tensor of small rank (r_1, r_2, r_3).

- Low rank tensor completion

$$\|\Omega \circ (\mathcal{T} - \mathcal{X})\|_F^2 + Regularization.$$

- For the low rank tensor completion problem, we use HOSVD and the approach in the time series case. For the Graph Laplacian Regularizer method, we unfold the tensor along the spatial mode to exploit the spatial relationship.

Low rank Tensor completion

- Matricize the tensor and apply matrix completion techniques -> we need to choose the right mode.
- We propose the following method: unfold the tensor along three modes. Along each mode, apply the matric completion techniques. Finally take the average of three solutions.
- We are still open to choose a setup/algorithm for this low rank tensor completion problem.

Thank You!