Modeling Multivariate Spatio-Temporal Remote Sensing Data with Large Gaps

Qiang Lou **Zoran Obradovic**
Center for Information Science and Technology, Temple University
Philadelphia, PA, USA
qianglou@temple.edu zoran@ist.temple.edu

Abstract

Prediction models for multivariate spatio-temporal functions in geosciences are typically developed using supervised learning from attributes collected by remote sensing instruments collocated with the outcome variable provided at sparsely located sites. In such collocated data there are often large temporal gaps due to missing attribute values at sites where outcome labels are available. Our objective is to develop more accurate spatio-temporal predictors by using enlarged collocated data obtained by imputing missing attributes at time and locations where outcome labels are available. The proposed method for large gaps estimation in space and time (called LarGEST) exploits temporal correlation of attributes, correlations among multiple attributes collected at the same time and space, and spatial correlations among attributes from multiple sites. LarGEST outperformed alternative methods in imputing up to 80% of randomly missing observations at a synthetic spatio-temporal signal and at a model of fluoride content in a water distribution system. LarGEST was also applied for imputing 80% of nonrandom missing values in data from one of the most challenging Earth science problems related to aerosol properties. Using such enlarged data a predictor of aerosol optical depth is developed that was much more accurate than predictors based on alternative imputation methods when tested rigorously over entire continental US in year 2005.

1 Introduction

Applicability of many existing data analysis methods that assume fairly complete data is limited by the presence of a large fraction of missing values (gaps) in data. These gaps are often large among spatio-temporal observations by remote sensing instruments due to the presence of clouds, malfunctions of remote sensing instruments and noise. Although many sound statistical methods exist for imputation of missing values (e.g. multiple imputation [Rubin *et al*, 1987]) most of existing methods are not applicable when gaps in data are very large. To address large gaps in surveys, data mining alternatives were proposed that exploit similarity information from shared neighbors [Ayuyev *et al*, 2009].

In multivariate spatio-temporal data analysis problems of natural systems in the presence of large gaps three kinds of correlation could be potentially exploited. These opportunities consist of temporal correlation of attributes, correlations among multiple attributes collected at the same time and space, and spatial correlations among attributes collected at multiple sites. Although many methods were proposed that take advantage of one or two of these aspects, less work was devoted to taking into consideration all three kinds of correlations simultaneously. Exploiting these correlations to impute large gaps in multivariate spatio-temporal data to develop more accurate supervised prediction models is the objective of our study.

Our work is motivated by the multivariate spatio-temporal prediction task of large interest in geosciences called Aerosol Optical Depth (AOD) retrieval. AOD retrieval data is obtained by integrating remote sensing observations with large gaps from multiple sources and with multiple kinds of correlation among attributes. In this application 19 real value satellite-based attributes were derived from multi-spectral images obtained once per day for the entire Earth by MODIS instrument on Terra satellite [Modis, 2011]. We use such satellite data provided at a 4km*4km grid covering the entire continental US in year 2005 (illustrated for two days at Figure 1 for a 36km*36km region). Although MODIS provides daily coverage of almost entire Earth, all 19 attributes are missing at nodes corresponding to locations where it was cloudy at time of the satellite overpass and such events are very common. Real valued AOD outcome variable is measured from ground at a small number of AERONET locations [Aeronet, 2011]. Due to high cost of such ground-based data collection, in year 2005 only 33 AERONET sites were available in the entire continental US. The objective of AOD retrieval is to use spatio-temporally collocated MODIS attributes and AERONET outcomes to build an accurate AOD predictor that can estimate AOD value from satellite attributes. In this application one of the main challenges is a small number of spatio-temporally collocated MODIS observations and AERONET AOD outcomes. In particular, in year 2005 for the entire Continental US there were only 805 spatio-temporal events where both satellite and ground-based data were available. Using the method proposed in this article,

This set is enlarged about five times to 4112 cases by imputing missing MODIS attributes by exploiting correlations among 19 attributes collected at a single node of the grid, spatial correlations of attributes observed at the same day at neighboring 80 nodes in 4x4km grid, and temporal correlations over multiple days. This allowed construction of much more accurate AOD retrievals as will be discussed in the results section.

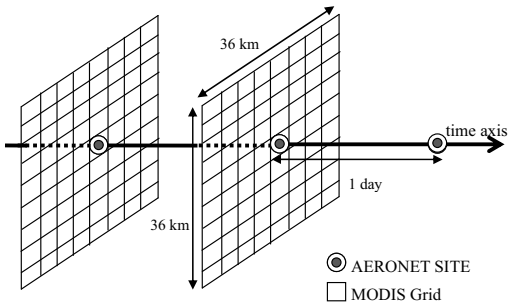

Figure 1. Integration of 19 MODIS attributes at 4km*4km daily grid with AOD outcome at a ground-based AERONET station.

The rest of the paper is organized as follows. The related work is briefly reviewed in Section 2. The proposed method for modeling multivariate spatio-temporal data is presented in Section 3 together with an algorithm for estimating missing observation. Evaluation results of the proposed method on both synthetic and real data sets are reported in Section 4. Section 5 provides the conclusion of our study.

2 Related Work

Many of the existing spatio-temporal data analysis methods assume that data is complete or almost complete. This assumption is violated in many spatio-temporal applications.

For estimation of missing values in multiple time series, a dynamic Bayesian model called DynaMMo was recently developed to simultaneously exploit temporal smoothness of each series and their correlations [Li *et al*, 2009]. Although very effective for estimating missing values in coevolving time series, DynaMMo is less applicable to remote sensing where data is collected at multiple spatially correlated locations where multiple correlated time series are observed at each location and individual series have temporal continuity.

For imputing incomplete spatial data, one of the most successful practical methods is to use multivariate interpolation by empirical orthogonal functions [Beckers *et al*, 2003]. In this singular value decomposition (SVD) based data imputation approach that we will simply call EOF, missing values are initially replaced by an unbiased guess and the missing values were interpolated incrementally by using truncated orthogonal functions of SVD decomposition for reconstruction and repeating the process while increasing the number of component functions. A limitation of EOF based data imputation is that it exploits only spatial correlations in data which is a problem when long continuous gaps are present in spatio-temporal data. An application of EOF to a transposed matrix that we will call T-EOF is proposed as a practical way to address such larger gaps in data [Kondrashov *et al*, 2006]. The same technique of using a transposed data to catch a different aspect of correlations in data (spatial instead of temporal) can also be applied to linear interpolation and to DynaMMo imputations. Such versions we will call T-Linear and T-DynaMMo methods.

Previous work [Radosavljevic *et al* 2010] in developing data-driven AOD retrieval methods using spatio-temporally collocated satellite and ground based observations (as shown at Figure 1) was simply removing the missing observations. The hypothesis explored in this study is that the accuracy of previously developed AOD predictors can be improved significantly by estimating the missing attributes and then train predictors on the data set consisting of both observed and imputed attributes.

3 Methodology

Given a multivariate spatio-temporal data (a multi-dimensional sequence and bunch of neighboring multi-dimensional sequences), there are three types of correlations: temporal correlation of each dimension, correlations among multiple dimensions collected at the same time and space, and spatial correlations from neighboring sites. Our goal is to build an accurate model on enlarged data with imputed values estimated by exploiting all three kinds of correlations.

3.1 Modeling correlations in a single sequence

In this section we describe how to model temporal correlation of each dimension and correlations among multiple dimensions from a multivariate sequence. We build a probabilistic model to estimate the missing values conditioned on observed values by exploiting these two types of correlation.

Assume that an m-dimensional sequence $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}\}$ of length N is given, where vector $\mathbf{x_n}$ observed at the n^{th} time tick of sequence (n = 1,…,N) is a m-dimensional multivariate Gaussian. For each m-dimensional observation of vector $\mathbf{x_n}$ we introduce a Gaussian latent variable $\mathbf{z_n}$ such that there is a linear dependence with a Gaussian noise between each $\mathbf{x_n}$ and $\mathbf{z_n}$ defined as $\mathbf{x_n} = \mathbf{C}\mathbf{z_n} + \mathbf{v_n}$, where $\mathbf{C}$ is the parameter matrix and $\mathbf{v} \sim N(\mathbf{v} \mid 0, \mathbf{\Sigma})$ is the Gaussian noise with mean of zero and variance of $\mathbf{\Sigma}$.

We also define a linear dependence with Gaussian noise between two adjacent latent variables $\mathbf{z_{n-1}}$ and $\mathbf{z_n}$ corresponding to two successive observation $\mathbf{x_{n-1}}$ and $\mathbf{x_n}$ as $\mathbf{z_n} = \mathbf{A}\mathbf{z_{n-1}} + \mathbf{w_n}$, where $\mathbf{A}$ is the parameter and $\mathbf{w} \sim N(\mathbf{w} \mid 0, \mathbf{\Gamma})$ is the Gaussian noise with mean of zero and variance of $\mathbf{\Gamma}$.

Therefore, the emission and transition distribution can be written as

$$p(\mathbf{x_n} \mid \mathbf{z_n}) = N(\mathbf{x_n} \mid \mathbf{C}\mathbf{z_n}, \mathbf{\Sigma}) \quad (1)$$

$$p(\mathbf{z_n} \mid \mathbf{z_{n-1}}) = N(\mathbf{z_n} \mid \mathbf{A}\mathbf{z_{n-1}}, \mathbf{\Gamma}) \quad (2)$$

The initial latent variable $\mathbf{z_1}$ also has a Gaussian distribution which can be written as

$$p(\mathbf{z_1}) = N(\mathbf{z_1} \mid \boldsymbol{\mu_0}, \mathbf{V_0}) \quad (3)$$

where μ_0 is the initial state of $\mathbf{z_1}$ and $\mathbf{V_0}$ is the variance.

Let $\boldsymbol{\theta} = \{\mathbf{A}, \boldsymbol{\Gamma}, \mathbf{C}, \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \mathbf{V}_0\}$ be the parameters of the model. Therefore, the joint probability given $\boldsymbol{\theta}$ is

$$p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}) = p(\mathbf{z_1} \mid \boldsymbol{\mu}_0, \mathbf{V}_0) \cdot \prod_{n=2}^{N} p(\mathbf{z_n} \mid \mathbf{z_{n-1}}, \mathbf{A}, \boldsymbol{\Gamma}) \cdot \prod_{n=1}^{N} p(\mathbf{x_n} \mid \mathbf{z_n}, \mathbf{C}, \boldsymbol{\Sigma}) \quad (4)$$

Our model is different to the traditional Kalman Filter -based model, since we allow missing values to exist in the observation $\mathbf{X}$. We define a Missing Index Matrix $\mathbf{I}$ to indicate the missing values. Each entry of $\mathbf{I}$ is defined as

$$\mathbf{I}_{pq} = \begin{cases} 0 & \text{when } \mathbf{X}_{pq} \text{ is missing} \\ 1 & \text{otherwise} \end{cases} \quad (5)$$

For learning the model, we define the expectation of the complete-data log likelihood as

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}_{old}) = E_{\mathbf{Z}|\boldsymbol{\theta}_{old}, \mathbf{I}}[\ln p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}, \mathbf{I})] \quad (6)$$

First, we initialize each missing value $\mathbf{X}_{pq}$ in data sequence (where $\mathbf{I}_{pq} = 0$) using simple linear interpolation from the values where $\mathbf{I}_{pq} \neq 0$ at the same time. Then, we apply the EM algorithm to maximize the equation (6). By extending the equation (6) by substituting $p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})$ using the corresponding part from equation (4) and (1)-(3), we get

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}_{old}) &= -\frac{1}{2}\ln|\mathbf{V_0}| - E_{\mathbf{Z}|\boldsymbol{\theta}_{old}}[\frac{1}{2}(\mathbf{z_1} - \boldsymbol{\mu}_0)^T \mathbf{V_0}^{-1}(\mathbf{z_1} - \boldsymbol{\mu}_0)] \\ &- \frac{N-1}{2}\ln|\boldsymbol{\Gamma}| - E_{\mathbf{Z}|\boldsymbol{\theta}_{old}}[\frac{1}{2}\sum_{n=2}^{N}(\mathbf{z_n} - \mathbf{A}\mathbf{z_{n-1}})^T \boldsymbol{\Gamma}^{-1}(\mathbf{z_n} - \mathbf{A}\mathbf{z_{n-1}})] \\ &- \frac{N}{2}\ln|\boldsymbol{\Sigma}| - E_{\mathbf{Z}|\boldsymbol{\theta}_{old}}[\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x_n} - \mathbf{C}\mathbf{x_n})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x_n} - \mathbf{C}\mathbf{x_n})] + const \end{aligned} \quad (7)$$

where *const* is the term which is not dependent on any part of parameter $\boldsymbol{\theta}$.We take the derivatives of equation (7) with respect to each part of parameter $\boldsymbol{\theta}$ and then set them to zero. We get the parameters updates as follows:

$$\boldsymbol{\mu}_0^{new} = E[\mathbf{z_1}] \quad (8)$$

$$\mathbf{V}_0^{new} = E[\mathbf{z_1}\mathbf{z_1}^T] - E[\mathbf{z_1}]E[\mathbf{z_1}^T] \quad (9)$$

$$\mathbf{A}^{new} = (\sum_{n=2}^{N} E[\mathbf{z_n}\mathbf{z_{n-1}}^T]) \cdot (\sum_{n=2}^{N} E[\mathbf{z_n}\mathbf{z_{n-1}}^T])^{-1} \quad (10)$$

$$\begin{aligned} \boldsymbol{\Gamma}^{new} = \frac{1}{N-1}\sum_{n=2}^{N}\{&E[\mathbf{z_n}\mathbf{z_n}^T] - \mathbf{A}^{new}E[\mathbf{z_{n-1}}\mathbf{z_n}^T] \\ &- E[\mathrm{B}^T]\mathrm{B}^{new} + \mathrm{B}^{new}E[\mathbf{z_{n-1}}\mathbf{z_{n-1}}^T](\mathbf{A}^{new})^T\} \end{aligned} \quad (11)$$

$$\mathbf{C}^{new} = (\sum_{n=1}^{N} \mathbf{x_n}E[\mathbf{z_n}^T]) \cdot (\sum_{n=1}^{N} E[\mathbf{z_n}\mathbf{z_n}^T])^{-1} \quad (12)$$

$$\begin{aligned} \boldsymbol{\Sigma}^{new} = \frac{1}{N}\sum_{n=1}^{N}\{&\mathbf{x_n}\mathbf{x_n} - \mathbf{C}^{new}E[\mathbf{z_n}]\mathbf{x_n}^T \\ &- \mathbf{x_n}E[\mathbf{z_n}^T]\mathbf{C}^{new} + \mathbf{C}^{new}E[\mathbf{z_n}\mathbf{z_n}^T](\mathbf{C}^{new})^T\} \end{aligned} \quad (13)$$

At the end of each M step, we update the missing value $\mathbf{X}_{pq}$ (when $\mathbf{I}_{pq} = 0$) using

$$E[\mathbf{X}_{pq} \mid \mathbf{Z}, \boldsymbol{\theta}, \mathbf{I}] = \mathbf{C}^{new} \cdot E[\mathbf{Z}_{\{p,q\}}] \ (when\ I_{pq} = 0) \quad (14)$$

Calculating the updated parameters requires the inference in E step of the marginal distribution $p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta})$ for hidden latent variables given the data. The inference is similarly to the one in Kalman Filter-based model, since the missing values are updated at the end of each M step. We apply forward-backward message passing to calculate the expectation of posterior distribution of latent variables. The details of inference in the Kalman Filter-based model are omitted for lack of space (for more details see [Bishop *et al*, 2006]).

Then, we use the updated $\mathbf{X}$ to recalculate the new parameters in the next EM iteration. We repeat this procedure until convergence. The estimation for missing values can be automatically obtained once the model is learned.

3.2 Modeling spatial correlations from neighbors

In this section, we describe how to explore spatial correlations among neighboring sites. We build a probabilistic model among multivariate sequence $\mathbf{X}$ and its neighboring observations to estimate the missing values in $\mathbf{X}$ conditioned on the observed values in neighboring sites.

Given a multivariate sequence $\mathbf{X}$, let $\mathbf{L_i} = \{l_{i1}, l_{i2}, \ldots, l_{iN}\}$ (i = 1,…m) be the ith dimension of $\mathbf{X}$, where l_{in} (n = 1,…,N) is a single value of the observation in the ith dimension at the nth time-step. Assuming that $\mathbf{X}$ has observations at *k* neighboring locations, we define $\mathbf{L}_i^{(j)} = \{l_{i1}^{(j)}, l_{i2}^{(j)}, ..., l_{iN}^{(j)}\}$ as the ith dimension of $\mathbf{X}$'s jth (j = 1,…,*k*) neighboring locations. Our objective is to impute the missing value in each $\mathbf{L_i}$ by exploiting the spatial correlation among $\mathbf{L_i}$ and $\{L_i^{(j)} \mid j = 1, ..., k\}$. In order to learn such spatial correlation, for each dimension $\mathbf{L_i}$ of $\mathbf{X}$ and corresponding $L_i^{(j)} = \{l_{i1}^{(j)}, l_{i2}^{(j)}, ..., l_{iN}^{(j)}\}$ from $\mathbf{X}$'s neighbors, we define $\mathbf{O}^{(i)} = \{\mathbf{o}_1^{(i)}, \mathbf{o}_2^{(i)}, ..., \mathbf{o}_N^{(i)}\}$ as the union of $\mathbf{L_i}$ and $\{\mathbf{L}_i^{(j)} \mid j = 1, ..., k\}$, where $\mathbf{o}_n^{(i)} = \{l_{in}, l_{in}^{(1)}, l_{in}^{(2)}, ...l_{in}^{(k)}\}$ (n = 1,…,N) are the values of the ith dimension of $\mathbf{X}$ and its neighboring locations at the nth time-step. For each observation $\mathbf{o}_n^{(i)}$, we define a Gaussian latent variable $\mathbf{y_n} \sim N(0, \mathbf{w})$ (n = 1,…,N). Each pair of nodes $\{\mathbf{y_n}, \mathbf{o_n}^{(i)}\}$ represents a linear-Gaussian latent variable model for the particular multivariate observation. However, the latent variables $\{\mathbf{y_n}\}$ are treated as independent to each other. Hence, the emission distribution is

$$p(\mathbf{o_n}^{(i)} \mid \mathbf{y_n}) = N(\mathbf{o_n}^{(i)} \mid \mathbf{D} \cdot \mathbf{y_n}, \phi) \quad (15)$$

Then, we can build a probabilistic graphical model for each dimension of $\mathbf{X}$ to exploit the spatial correlation between each $\mathbf{L_i}$ and its corresponding $\{\mathbf{L}_i^{(j)} \mid j = 1, ..., k\}$ from neighboring locations. Let $\boldsymbol{\gamma} = \{\mathbf{w}, \mathbf{D}, \phi\}$ be the parameters of the model. Then, the joint distribution can be written as:

$$p(\mathbf{O}^{(i)}, \mathbf{Y} \mid \boldsymbol{\gamma}) = \prod_{n=1}^{N} p(\mathbf{y_n}) \cdot \prod_{n=1}^{N} p(\mathbf{o_n}^{(i)} \mid \mathbf{y_n}, \boldsymbol{\gamma}) \quad (16)$$

Therefore, maximizing the complete data log likelihood is equivalent to maximizing:

$$\begin{aligned} \ln(\prod_{n=1}^{N} p(\mathbf{y_n}) \prod_{n=1}^{N} p(\mathbf{o_n}^{(i)} \mid \mathbf{y_n}, \boldsymbol{\gamma})) &= -\frac{N}{2}\ln|\mathbf{w}| - \sum_{n=1}^{N}\frac{1}{2}\mathbf{y_n}^T\mathbf{w}^{-1}\mathbf{y_n} \\ &- \frac{N}{2}\ln|\phi| - \sum_{n=1}^{N}\frac{1}{2}(\mathbf{o_n}^{(i)} - \mathbf{D} \cdot \mathbf{y_n})^T\phi^{-1}(\mathbf{o_n}^{(i)} - \mathbf{D} \cdot \mathbf{y_n}) \end{aligned} \quad (17)$$

We take the derivatives of equation (17) with respect to $\mathbf{W}$, $\mathbf{D}$ and ϕ respectively, and set them to zero. The updated parameters are computed as

$$\mathbf{w}^{new} = \frac{1}{N}\sum_{n=1}^{N} \mathrm{E}(\mathbf{y_n} \mid \mathbf{o_n}^{(i)})\mathrm{E}(\mathbf{y_n}^T \mid \mathbf{o_n}^{(i)}) \quad (18)$$

$$\mathbf{D}^{\text{new}} = \sum_{n=1}^{N} \mathbf{o_n}^{(i)} E(\mathbf{y_n}^T \mid \mathbf{o_n}^{(i)}) \cdot (\sum_{n=1}^{N} E(\mathbf{y_n} \mid \mathbf{o_n}^{(i)}) E(\mathbf{y_n}^T \mid \mathbf{o_n}^{(i)}))^{-1} \quad (19)$$

$$\phi^{new} = \frac{1}{N}\sum_{n=1}^{N} (\mathbf{o_n}^{(i)} - \mathbf{D}^{new} \cdot E(\mathbf{y_n} \mid \mathbf{o_n}^{(i)})) \cdot (\mathbf{o_n} - \mathbf{D}^{new} \cdot E(\mathbf{y_n} \mid \mathbf{o_n}^{(i)}))^T \quad (20)$$

In order to get the optimal parameters maximizing equation (17) in the presence of missing observations, for each dimension of $\mathbf{X}$ we maintain another Missing Index Matrix $\mathbf{I}^{(i)}$ where $\mathbf{I}^{(i)}_{pq} = 0$ indicates a missing value of $\mathbf{O}^{(i)}_{pq}$. We initialize the each missing value using linear interpolation from values where $\mathbf{I}^{(i)}_{pq} \neq 0$ in the neighboring observations. Then, we calculate new parameters using equation (18)-(20) and use them to estimate the missing values as

$$E[\mathbf{O}_{pq(new)}^{(i)} \mid \mathbf{Y}, \boldsymbol{\gamma}, \mathbf{I}^{(i)}] = \mathbf{D}^{new} \cdot E[\mathbf{Y}_{\{p,q\}} \mid \mathbf{O}_{pq}^{(i)}] \ (when\ \mathbf{I}^{(i)}_{pq} = 0) \quad (21)$$

After imputing the missing values using eq. (21), we use the new data to estimate new parameters in the next iteration. By repeating this process of estimating parameters and missing values until converging, we can get the optimal parameters of the model and the final estimation of missing values. After updating the missing values in $\mathbf{O}^{(i)}_{pq}$ for each dimension of $\mathbf{X}$, we can get the estimated $\mathbf{X}$.

3.3 Learning algorithm

In order to estimate the missing values by exploring all three types of correlation, we propose the LarGEST algorithm (shown in Algorithm 1) which simultaneously learns two models described in Sections 3.1 and 3.2.

First, we initialize all model parameters and fill all the missing values by linear interpolation from the values of spatial neighbors. Then, we apply an Extended Expectation Maximization algorithm which works as follows.

In the E-step, we estimate the posterior distribution $p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta})$ of Model 1 which will be used when we maximize the expectation of log likelihood in M-step (using equations (8) -- (13)). After getting the updated parameters of Model 1, we can estimate the missing values using Model 1. The data with updated missing values from Model 1 is used to estimate the parameters of Model 2. We can re-estimate the missing values after learning the parameters of Model 2. The updated data with updated missing values estimated by Model 2 will be the input data of E-step of next iteration to calculate the posterior distribution of Model 1. We repeat this procedure of training two models interactively until convergence. Our experiments show that different order of two models has no significant influence in learning results since it only results in the different initialization values of two models.

After imputing the missing values in multivariate spatio-temporal data, we then build a predictor on enlarged collocated spatio-temporal data. In the next section we compare results of predictors trained on enlarged dataset generated by LarGEST and by alternative methods.

Inputs: $\mathbf{X}$ – set of multivariate spatio-temporal sequences with missing values
$\{\mathbf{O}^{(i)}\}$ – multivariate observations including neighbors for each dimension of $\mathbf{X}$
$\mathbf{I}$ – Missing value Index Matrix for $\mathbf{X}$
$\mathbf{I}^{(i)}$ –Missing value Index Matrix for $\mathbf{O}^{(i)}$
Outputs: $\mathbf{X}^{\text{new}}$ – sequences with estimated values
$\boldsymbol{\theta}$, $\boldsymbol{\gamma}$ – model parameters

Initialize $\mathbf{X}^{\text{new}}$ with $\mathbf{X}$
Estimate missing values in $\mathbf{X}^{\text{new}}$ using linear interpolation
Initialize model parameters
Do
  E-step: Estimate posterior distribution $p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta})$ of Model 1 using forward-backward message passing.
  M-step: Maximize expectation of log likelihood (Model 1)
    $\boldsymbol{\theta}^{new} \leftarrow \arg\max Q(\boldsymbol{\theta})$
  Estimate missing values using Model 1:
    for p, q DO
      update $\mathbf{X}^{\text{new}}{}_{pq}$ when $\mathbf{I}_{pq}$=0 using equation (14)
  Initialize $\{\mathbf{O}^{(i)}_{new}\}$ **with** $\mathbf{X}^{\text{new}}$ **and** $\{\mathbf{O}^{(i)}\}$
  Maximize log-likelihood (Model 2 for each dimension i):
    $\boldsymbol{\gamma}^{new} \leftarrow \arg\max \ln p(\mathbf{O}^{(i)}_{new}, \mathbf{Y} \mid \boldsymbol{\gamma})$
  Estimate missing values using Model 2
    for p, q DO
      update $\mathbf{O}_{pq(new)}^{(i)}$ when $\boldsymbol{I}_{pq}$=0 using eq. (21)
    $\mathbf{X}^{\text{new}} = \{\mathbf{O}^{(i)}_{new}\}$
Until converge
Return $\mathbf{X}^{\text{new}}$, $\boldsymbol{\theta}$, and $\boldsymbol{\gamma}$

Algorithm 1. LarGEST algorithm.

4 Experimental Results

The proposed LarGEST method for imputation of large gaps in data is compared experimentally with Linear, DynaMMO and EOF methods. In addition, LarGEST is also compared to the same three methods when used on transposed data and such results are annotated as T-Linear, T-DynaMMo and T-EOF. Brief descriptions of six alternative methods are provided in Section 2.

Our evaluation is performed on three datasets of increasing complexity. In the first task (Section 4.1) the objective was to compare LarGEST to alternatives for imputing 5% to 90% missing values in a fairly simple function. In Section 4.2 this is followed by imputing 80% of missing values in a more challenging problem of Floride estimation whereas a similar data was used as a testbed in an earlier study [Li et al, 2009]. Finally, in Section 4.3 effects of data imputation by LarGEST and alternatives were compared at a grant challenge problem of spatio-temporal regression for Aerosol AOD retrieval from data with about 80% of missing values in 19 attributes.

4.1 Imputation of small versus large gaps

The synthetic two dimensional temporal data (X, Y) for a certain location is generated as $X = \sin(t) * 5$ and $Y = \sin(t + \pi/2) * 5$. At four neighboring locations two dimensional data is generated by shifting the first site data by 0.5 0.3, -0.3and -0.5 respectively.

The mean square error (MSE) of imputation by LarGEST and six alternatives was compared when 5% to 90% of data were missing randomly (see Figure 2). In data imputation experiments with small fraction (less than 25%) of missing values inserted completely at random, LarGEST and all alternative methods estimated missing values fairly well. However, LarGEST was clearly the best choice as Linear interpolation had problems when the missing values were located near the top or the bottom of the signal, while EOF had problems where four neighbors were missing and DynaMMo had some errors at high curvature sections.

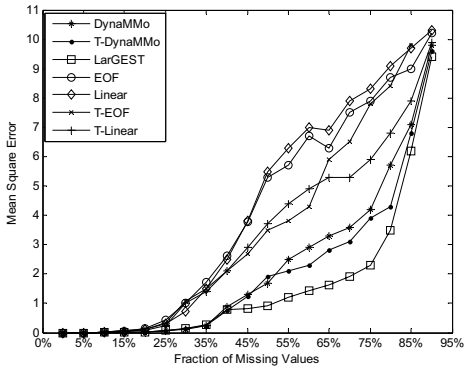

Figure 2. Estimated error (MSE) for imputing 5% to 90% of missing values on synthetic data.

When large fraction of data was missing LarGEST exploited well all three kinds of correlations in data simultaneously while alternative imputation methods resulted in much larger error. When extremely large fraction (85% and 90%) of data was missing, all methods performed badly. For a large fraction of missing values accuracy was improved when the number of neighbors was increased. Results with a larger number of neighbors are omitted on synthetic data for lack of space, but will be shown on real remote sensing data.

4.2 Imputation of 80% gaps in Fluoride data

The Fluoride dataset is produced by EPANET 2 which models the water quality behavior in a distribution piping systems. In a given network of water distribution piping system, EPANET simulates the contents of Fluoride over a certain period [EPA, 2011]. Data used in our experiments are generated assuming a piping system of 36 nodes simulated over 10 days in 15 minutes increments. For our experiments at a certain node and at its three neighboring nodes we extracted two attributes over 960 time steps. Then, 80% of this data was removed completely at random from each of two attributes at each of 4 nodes retaining 192 values per each attribute at each site. The objective was to estimate 1536 missing values at two time series observed at one of these nodes. The mean square error (MSE) of this estimation by LarGEST and alternative methods are shown at Figure 3.

4.3 Spatio-temporal regression of Aerosol data with 80% of imputed attributes

Experiments were conducted on integrated satellite and ground based Aerosol data introduced in Section 1. AOD

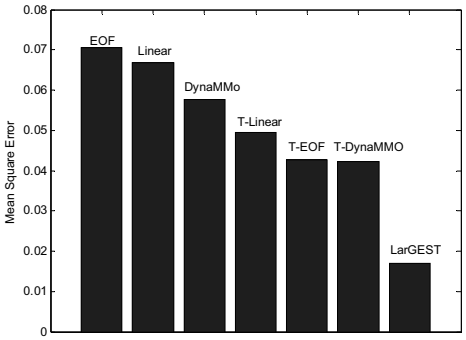

Figure 3. Mean square error of data imputation on Fluoride data with **80%** of missing values.

values observed from ground at 33 AERONET sites were also available at an additional 3,307 events, but in these cases all 19 sattelite-based attributes were missing. Missing attributes at these 80% cases were imputed by LarGEST relaying on spatial correlations with satellite observations at up to 80 neighbors at 4kmx4km grid as well as on temporal correlations among 330 daily observations. Six alternative methods used in Section 4.1 were also applied for attributes imputation at 3,307 events.

A feed-forward neural network model with a single hidden layer of 10 nodes was trained on enlarged data consisting of examples with actual and those with imputed attributes. This choice was based on the best predictors from previous studies [Radosavljevic. *et al*, 2010]. Experiments were performed by partitioning 805 examples from 33 sites where both attributes and AOD values are available in 33 disjoint subsets based on sites and using 32 subsets together with 3,307 additional examples whose attributes are imputed for training a neural network model which is tested on the remaining site's data. This is repeated in 33-cross validation experiments always keeping a different site for testing. The quality of the obtained predictors was compared using two measures following a protocol practiced by geoscience community [Radosavljevic *et al*, 2010]. To evaluate impact of spatial neighborhood size on imputation quality, in LarGEST imputation we considering nearest 8, 24, 48 and 80 neighboring observations in 4kmx4km grid shown at Figure 1. These methods we will call LarGEST8, LarGEST24, LarGEST48 and LarGEST80, respectively.

Learning time is not a concern in this application. In 10 minutes our method imputes data streams used in our experiments on a desktop computer with only 3 GB memory. This time is negligible as compared to months of simulations on a supercomputer required by traditional deterministic aerosol retrieval methods based on physical modeling.

The first quality measure we used is R-square. The results obtained by a predictor trained on data imputed by LarGEST80, and by nine alternative methods are shown in Figure 4. In addition, we also show the accuracy of a predictor obtained on 805 examples without data imputation and we call this model Original. The results obtained by LarGEST80 were more accurate than alternatives and better than any previously reported accuracy of AOD retrieval.

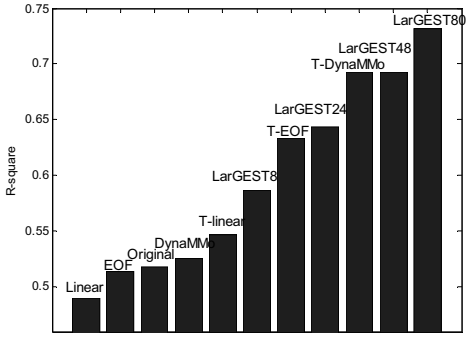

Figure 4. R-square accuracy of AOD retrieval in US for year 2005using data with 80% of imputed values

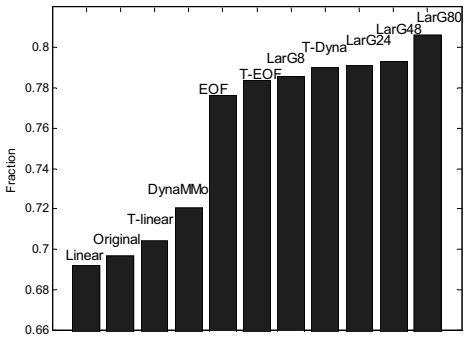

Figure 5. Fraction of successful AOD retrievals in US for year 2005 when using 80% imputed values (LarG in the figure means LarGEST)

Another measure we used is introduced by aerosol scientists who regard the retrieved AOD acceptable if the following boundary conditions on the retrieval are satisfied $| y_i - t_i | \leq 0.05 + 0.15 t_i$ where y_i is retrieved AOD value and t_i is the corresponding true AOD value. We measured fraction of successful predictions defined as $FRAC = (I / N) * 100\%$, where I is the number of predictions satisfying the stated boundary and N is the total number of predictions. Fraction of successful AOD retrievals of seven regression models based on data imputed by LarGEST and six alternatives as well as data that is not imputed at all (called Original) are shown at Figure 5. LarGEST80 clearly outperformed all alternative methods in both measures.

Both Figure 4 and Figure 5 show that LarGEST performed better when including attributes from larger number of neighboring grids. All 80 neighbors used in LarGEST80 model were within the range of 36km (see Figure 1) which is considered to be a range of AOD spatial correlation. Indeed, including spatial information from even more distant nodes was not beneficial (results omitted for the lack of space).

5 Conclusion

In the proposed method, two probabilistic models were proposed, and learned interactively by an Extended Expectation Maximization algorithm to exploit simultaneously all three types of correlation for multivariate spatio-temporal data imputation. The imputation results on challenging problems with 80% of missing values provide evidence that in the presence of long continuous gaps LarGEST method can estimate missing values more accurately than alternatives. The aerosol optical depth retrieval results obtained using training data enlarged by LarGEST-based imputations were not only better than the results obtained by training a predictor trained on data imputed using alternative methods, but were also more accurate than any previously developed method for AOD retrieval from satellite data.

Acknowledgments

This work was supported in part by the National Science Foundation under Grant IIS-0612149.

References

[Ayuyev *et al*, 2009] Ayuyev, V., Jupin, J., Harris, P., Obradovic, Z. "Dynamic clustering based estimation of missing values in mixed type data," Proc. 11th Int'l Conf. Data Warehousing and Knowledge Discovery, Linz, Austria, 2009, pp. 366-377.

[Beckers *et al*, 2003] Beckers, J. and Rixen, M., "EOF calculations and data filling from incomplete oceanographic data sets," J. Atmos. Ocean. Technol., vol. 20, 2003, pp. 1839–1856,.

[Bishop *et al*, 2006] Bishop, C. M., "Pattern Recognition and Machine Learning," Springer, Aug. 2006.

[Aeronet, 2011] http://aeronet.gsfc.nasa.gov/new_web

[Modis, 2011] http://modis.gsfc.nasa.gov

[EPA, 2011] http://www.epa.gov/nrmrl/wswrd/dw/

[Khan *et al*, 2009] Khan, R.A., Nelson, D.L., Garay, M.J., Levy, R.C., Bull, M.A., Diner, D.J., Martonchik, J.V., Paradise, S.R., Hansen, E.G., Remer, L.A.,"MISR aerosol product attributes and statistical comparisons with MODIS," IEEE Tran. Geoscience and Remote Sensing, vol. 47(2), no, 12, 2009, pp. 4095-4114.,.

[Kondrashov *et al*, 2006] Kondrashov, .D and Ghil, M, "Spatio-temporal filling of missing points in geophysical data sets," Nonlinear Processes Geophys, 13, 2006, pp. 151-159.

[Li *et al*, 2009] Li, L., McCann, J., Pollard, N., Faloutsos,C. "DynaMMo: Mining and summarization of Coevolving Sequences with missing values," Proc. 15th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, , Paris, France, 2009, pp. 507-516.

[Radosavljevic *et al*, 2010] Radosavljevic, V., Vucetic, S., Obradovic, Z. "A data mining technique for aerosol retrieval across multiple accuracy measures," IEEE Geoscience and Remote Sensing Letters, vol. 7, no. 2, 2010, pp. 411-415.

[Rubin *et al*, 1987] Rubin, D. B. , "Multiple imputation for nonresponse in surveys," New York:Wiley, 1987.