Robust Prediction of Large Spatio-Temporal Datasets

Yang Chen

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Science and Applications

Ing-Ray Chen, Chair
Charles Clancy, Co-Chair
Guoqiang Yu

May 23, 2013
Falls Church, Virginia

Keywords: Robust Prediction, Expectation Propagation, Student's t Model,
Bayesian Hierarchical Model, Spatio-Temporal Process

Robust Prediction of Large Spatio-Temporal Datasets

Yang Chen

ABSTRACT

This thesis describes a robust and efficient design of Student-t based Robust Spatio-Temporal Prediction, namely, St-RSTP, to provide estimation based on observations over spatio-temporal neighbors. It is crucial to many applications in geographical information systems, medical imaging, urban planning, economy study, and climate forecasting. The proposed St-RSTP is more resilient to outliers or other small departures from model assumptions than its ancestor, the Spatio-Temporal Random Effects (STRE) model. STRE is a statistical model with linear order complexity for processing large scale spatiotemporal data.

However, STRE has been shown sensitive to outliers or anomaly observations. In our design, the St-RSTP model assumes that the measurement error follows Student's t-distribution, instead of a traditional Gaussian distribution. To handle the analytical intractable inference of Student's t model, we propose an approximate inference algorithm in the framework of Expectation Propagation (EP). Extensive experimental evaluations, based on both simulation and real-life data sets, demonstrated the robustness and the efficiency of our Student-t prediction model compared with the STRE model.

Acknowledgments

I would like to express my gratitude to my advisor T. Charles Clancy for giving me the freedom to satisfy my curiosity, not only in the research but many others. At the time when I enrolled in the HUME center, he started to give his big help to support us. Even in the hardest time of my life, he let me go back to visit my mother for long time and continuously expressed his concerned about my families.

I would like to give my respect and thankful to Feng Chen for working with me, helping me in every aspect of life. He led me to deeply interest in machine learning, data mining and encouraged me to conquer the hard machine learning problems by probability theory. For comments on drafts of this thesis, I gave my deepest appreciation to my committees Ing-Ray Chen, Guoqiang Yu and advisor. The department head Ing-Ray Chen gave me a lot suggestion about the research and thesis work.

I also would like to thank the City of Bellevue, Washington for providing arterial traffic data. I am also grateful to the STAR Lab for maintaining the arterial database and providing the online portal to access the data.

Finally, I would like to thank my family member, especially my parents. They raised me up, taught me to be a responsible person, encouraged me to study abroad, even though they missed me so much. I also thank all of my friends for their help.

Contents

1 Introduction **1**

1.1 Motivation 2

1.2 Contribution 3

2 Theoretical Backgrounds & Related Works **4**

2.1 Spatio-Temporal Random Effects Model 4

2.2 The Forward-Backward Algorithm 7

2.2.1 Forward Recursion 8

2.2.2 Backward Recursion 10

2.3 Expectation Propagation 12

3 Student-t Robust Spatio-Temporal Prediction Model **15**

3.1 Student-t RSTP Model 15

3.2 Approximating $p(\eta_t|\mathbf{Z}_{1:t})$ and $p(\eta_t|\mathbf{Z}_{1:T})$ 17

3.3 Approximating $p(\xi_t|\mathbf{Z}_{1:t})$ and $p(\xi_t|\mathbf{Z}_{1:T})$ 22

4 Experiments **27**

4.1 Experiment Design and Evaluation 27

4.1.1 Experiment Design 27

4.1.2 Result Evaluation 28

4.2 Simulation Study 29

4.2.1 Simulation Setup 29

4.2.2 Simulation Results 30
4.3 Aerosol Optical Depth Data Experiments 33
4.4 Case Study on Traffic Volume Data 35
4.5 Time Complexity 37

5 Conclusion **39**

List of Abbreviations

AOD	Aerosol Optical Depth
EM	Expectation Maximization
EP	Expectation Propagation
IRLS	Iterative Reweighted Least Squares
MAPE	Mean Absolute Percentage Error
MCMC	Markov Chain Monte Carlo
RMSE	Root Mean Square Error
STRE	Spatio-Temporal Random Effect
St-RSTP	Student-t Robust Spatio-Temporal Prediction
TV	Traffic Volume

List of Figures

2.1 Illustration of forward recursion 9

2.2 Illustration of backward recursion 11

2.3 Intuition of Expectation Propagation 13

3.1 Student's t vs. Gaussian Distribution 16

3.2 St-RSTP Graphic Model 17

3.3 Factor Graph Presentation of St-STRE 19

4.1 Experiment Design 28

4.2 STRE filtering vs. St-RSTP filtering using simulation data (X-axis: location, Y-axis: Z or Y value) 31

4.3 STRE vs. St-RSTP on AOD data sets at time unit 5 34

4.4 STRE vs. St-RSTP using the TV data on 5th day 36

4.5 Time Cost vs. Number of Locations 38

List of Tables

4.1 Model Robustness Comparison using Different Simulation Settings 33
4.2 Model Robustness Comparison use the AOD Data 35
4.3 Comparison of Time Cost using the Simulated and AOD Data 37

Chapter 1

Introduction

Spatial and temporal information exist almost everywhere in the real world. Most physical and biological processes involve spatial and temporal variability to some extent [[2], [22], [7]]. For applications such as geographical information systems, medical imaging, urban traffic prediction, and weather forecasting, considering spatial and temporal information helps separate causalities due to pure environmental effects. Modeling spatial temporal processes has been regarded as an essential component in many environmental modeling systems. It is suggested that any application that involves stochastic process as a component should consider spatial and temporal dependencies [5].

There have been two paradigms for modeling spatio-temporal data, including Kriging based and dynamical (mechanic or probabilistic) specification based. The Kriging based paradigm basically extends spatial dimensions (d) with an extra time dimension and focuses on the modeling of the variance-covariance structure between the observations in the $(d+1)$-dimensional space. Different joint time-space covariance structures have been proposed to model the heterogeneities between temporal and spatial dimensions based on different scenarios. The dynamic specification based paradigm considers spatio-temporal processes through a dynamical-statistical (or state space based) framework. The observations in the current state are dependent on its previous states through dynamic mechanical (or probabilistic) relationships. This thesis focuses on the dynamic statistical paradigm, which can be explicitly specified based on the knowledge of the phenomenon under study, always leads to a valid variance-covariance structure, and allows fast filtering, smoothing, and forecasting. These advantages have been well validated in the time series and signal processing literatures (e.g., [[12], [1], [25]]).

1.1 Motivation

Currently, one emerging research challenge is to design efficient algorithms to process massive spatio-temporal data that have been collected by using advanced remote sensing technology. For example, National Aeronautics and Space Administration (NASA) has launched satellites (e.g., Terra satellite) that have the ability to collect data on the order of 100,000 observations per day. Given the large volume of remote sensing data, most traditional spatio-temporal statistical models fail to process in either allowable memory space limit or an acceptable time limit, even in supercomputing environments. [10] presents spatio-temporal kalman filters, with the related works followed by [18], [27], and others. [6] presents a summary of the related literature. [11] presents a multi-resolution filtering algorithm that basically considers blocky basis functions and coarse resolution dynamics. [8] presents a Bayesian framework that decomposes the variance-covariance matrix into two components, including an upper triangular matrix and a diagonal matrix. [16] presents a Bayesian spatial dynamic factor-analysis model. Although this model improves the computational efficiency, it still has a high model complexity with a number of parameters and requires Markov Chain Monte Carlo (MCMC) for conducting inferences, which is still computationally expensive. [3] presents a full Bayesian spatio-temporal model to handle irregularly sampled data and achieves numerical model output, but it still requires MCMC to do inferences.

Although progresses have been made, all the preceding works are still unable to achieve near-real-time performance and thus not suitable for processing massive and streaming spatial data. As the most recent advancement, [5] presents a spatio-temporal random effects (STRE) model that reduces the problem into a fixed dimension problem and makes it possible to do fast filtering, smoothing, and prediction with a linear order time complexity. The importance of fast filtering and smoothing techniques lies in the fact that these techniques can be viewed as a type of data assimilation and be used to re-initialize a numerical forecast. As reviewed in the recent article by [26], "... (data assimilation) is an approach for fusing data (observations) with prior knowledge (e.g., mathematical representations of physical laws; model output) to obtain an estimate of the distribution of the true state of a process."

The STRE model achieves a good performance with a linear order time complexity. However, this model does not allow a latent component to explicitly model the variations due to outliers or small departures from model assumptions. Studies have also shown that the prediction results based on STRE are clearly distorted when a small portion of outliers are involved. This observation motivates us to propose a robust version of the STRE model and to design efficient inference algorithms that have comparable time costs.

1.2 Contribution

In this work, we present the Student-t observation model for Spatio-temporal prediction because of its good robustness properties which can be altered continuously from a very heavy tailed distribution to the Gaussian model with the degrees of freedom parameter. However, the challenge with the Student-t model is the analytically intractable inference. A number of methods can be applied here, such as Gibbs sampling, factorizing varitional approximation, and expectation propagation. A a most recent work by Pasi Jylanki, the Expectation Propagation framework shows the most favorable results on both efficiency and effectiveness for robust gaussian process regression. Here, in our model, it is basically a reduced rank gaussian process coupled with a lag-1 temporal autoregressive model. We propose an robust implementation for the prediction inference in the framework of Expectation Propagation. We show that, our method outperformed the standard STRE via a simulation study and experiments on a number of real data sets.

The main contributions of our study can be summarized as follows.

- We formalize an innovative robust predictive process for spatio-temporal data in a systematical framework;
- We approximate a robust prediction model such that the high-dimensional latent variables can be separated into groups that can be optimized iteratively;
- We present novel implementations of Expectation Propagation (EP) in order to efficiently estimate the posterior distributions of latent variables.
- We validate the robustness and the efficiency of the proposed St-RSTP model compared with the regular STRE model by an extensive simulation study and experiments on two real data sets.

The rest of the paper is organized as follows. Preliminaries on the formulation and inference algorithms of the regular STRE model and the EP framework are reviewed in Chapter 2. Chapter 3 presents the robust spatio-temporal prediction model, St-RSTP, followed by the detailed prediction techniques based on EP. Simulation study and evaluation of our proposed robust smoothing algorithm on two real world data sets are illustrated in Chapter 4. Finally, we conclude our work in Chapter 5.

Chapter 2

Theoretical Backgrounds & Related Works

This Chapter first reviews the Spatio-Temporal Random Effects model, including the model formulation and the filtering and smoothing process. Before the intractable solver techniques, Expectation Propagation, it shows the forward-backward algorithm using in linear dynamic system, which is a very popular technique in message passing. It then describes the approximation approach, Expectation Propagation, which has been efficient and effective in a number of studies for approximate inferences [23].

2.1 Spatio-Temporal Random Effects Model

Many data sets contain spatial and temporal information attached to the attribute information. And closer observations in space or time generally result in high statistical correlation. This dependence can be described through specification of a spatio-temporal covariance function, or it can be explained through a dynamical model that gives either a probabilistic or a statistical-physical mechanism for the evolution of the present and past. The Spatio-Temporal Random Effects (STRE) model is a statistical model proposed recently for processing large spatio-temporal data in linear order time complexity [5]. The filtering, smoothing, and forecasting based on STRE are also named fixed rank filtering, smoothing, and forecasting. The STRE model is used to model a spatial random process that evolves over time, $\{Y_t(\mathbf{s}) \in \Re : \mathbf{s} \in D \subset \Re^2, t = 1, 2, \cdots\}$, where D is the spatial domain under study, and $Y_t(\mathbf{s})$ is the nonspatial measurement (e.g., temperature) at location $\mathbf{s}$ and time t.

A discretized version of the process can be represented as

$$\{\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_t, \mathbf{Y}_{t+1}, \cdots\}, \tag{2.1}$$

where $\mathbf{Y}_t = [Y_t(\mathbf{s}_{1,t}), Y_t(\mathbf{s}_{2,t}), \cdots, Y_t(\mathbf{s}_{m_t,t})]^T$. The sample locations $\{\mathbf{s}_{1,t}, \mathbf{s}_{2,t}, \cdots, \mathbf{s}_{m_t,t}\}$ can

be different spatial locations at different time t.

Two major uncertainties, including missing data and noise (measurement error) can be handled in this model. Suppose we have the measurements $\{\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_t, \mathbf{Z}_{t+1}, \cdots\}$, with

$$\mathbf{Z}_t = \mathbf{O}_t \mathbf{Y}_t + \varepsilon_t, t = 1, 2, \cdots, \tag{2.2}$$

where $\mathbf{Z}_t$ is an n_t-dimensional vector ($n_t \leq m_t$), $\mathbf{O}_t$ is an $n_t \times m_t$ incidence matrix, which presents the relation of presenting locations at time t to all possible locations, and $\varepsilon_t = [\varepsilon_t(\mathbf{s}_{1,t}), \cdots, \varepsilon_t(\mathbf{s}_{n_t,t})]^T \sim \mathcal{N}_{n_t}(\mathbf{0}, \sigma^2_{\varepsilon,t}\mathbf{V}_{\varepsilon,t})$ is a vector of white noise Gaussian processes, with $\mathbf{V}_{\varepsilon,t} = diag(v_{\varepsilon,t}(\mathbf{s}_{1,t}), \cdots, v_{\varepsilon,t}(\mathbf{s}_{n_t,t}))$. Particularly, $var(\varepsilon_t(\mathbf{s})) = \sigma^2_{\varepsilon,t} v(\mathbf{s}) > 0, \sigma^2_{\varepsilon,t}$ is a parameter to be estimated, and $v(\mathbf{s})$ is known. The white noise assumption implies that $cov(\varepsilon_t(\mathbf{s}), \varepsilon_u(\mathbf{r})) = 0$, for $t \neq u$ and $\mathbf{s} \neq \mathbf{r}$.

Assume that $\mathbf{Y}_t$ has the following structure:

$$\mathbf{Y}_t = \mathbf{X}_t \beta_t + \nu_t, t = 1, 2, \cdots, \tag{2.3}$$

where $\mathbf{X}_t\beta_t$ is a deterministic (spatio-temporal) mean function, or trend, modeling large-scale variation. $\mathbf{X}_t = [\mathbf{x}_t(\mathbf{s}_{1,t}), \cdots, \mathbf{x}_t(\mathbf{s}_{m_t,t})]^T$, $\mathbf{x}_t(\mathbf{s}_{i,t}) \in \Re^p, 1 \leq i \leq m_t$, represents a vector of covariates, and the coefficients $\beta_t = (\beta_{1,t}, \cdots, \beta_{p,t})^T$ are general unknown. The random process ν_t captures the small scale variations. For traditional spatio-temporal Kalman filtering models, a large number of parameters need to be estimated with high computational costs due to high data dimensionality during the filtering, smoothing, and prediction processes. As a key advantage of the STRE model, it models the small scale variation ν_t as a vector of spatial random effects (SRE) processes

$$\nu_t = \mathbf{S}_t^T \eta_t + \xi_t, t = 1, 2, \cdots, \tag{2.4}$$

where $\mathbf{S}_t = [S_t(\mathbf{s}_{1,t}), \cdots, S_t(\mathbf{s}_{m_t,t})]$, $S_t(\mathbf{s}_{i,t}) = [S_{1,t}(\mathbf{s}_{i,t}), \cdots, S_{r,t}(\mathbf{s}_{i,t})]^T$, $1 \leq i \leq m_t$, is a vector of r predefined spatial basis functions, such as wavelet and bisquare basis functions, and η_t is an r-dimensional zero-mean Gaussian random vector with an $r \times r$ covariance matrix given by $\mathbf{K}_t$. The first component in Equation (2.4) denotes a smoothed small-scale variation at time t, captured by the set of basis functions $\{S_t(\mathbf{s}_{1,t}), \cdots, S_t(\mathbf{s}_{m_t,t})\}$.

The second component in Equation (2.4) captures the micro-scale variability similar to the nugget effect as defined in geostatistics [5]. It is assumed that $\xi_t \sim \mathcal{N}_{m_t}(\mathbf{0}, \sigma^2_{\xi,t}\mathbf{V}_{\xi,t})$, $\mathbf{V}_{\xi,t} = diag(v_{\xi,t}(\mathbf{s}_{1,t}), \cdots, v_{\xi,t}(\mathbf{s}_{m_t,t}))$, and $v_{\xi,t}(\cdot)$ describes the variance of the micro-scale variation and is typically considered known. Note that the component ξ_t is important, since it can be used to capture the extra uncertainty due to the dimension reduction in replacing ν_t by $\mathbf{S}_t^T\eta_t$. The coefficient vector η_t is assumed to follow a vector-autoregressive process of order one,

$$\eta_t = \mathbf{H}_t \eta_{t-1} + \zeta_t, t = 1, 2, \cdots, \tag{2.5}$$

where $\mathbf{H}_t$ refers to the so-called propagator matrix, $\zeta_t \sim \mathcal{N}(0, \mathbf{U}_t)$ is an r-dimensional innovation vector, and $\mathbf{U}_t$ is named as the innovation matrix. The initial state $\eta_0 \sim \mathcal{N}_r(\mathbf{0}, \mathbf{K}_0)$ and $\mathbf{K}_0$ is in general unknown.

Combining Equations (2.2), (2.3), and (2.4), the (discretized) data process can be represented as

$$\mathbf{Z}_t = \mathbf{O}_t\mu_t + \mathbf{O}_t\mathbf{S}_t^T\eta_t + \mathbf{O}_t\xi_t + \varepsilon_t, t = 1, 2, \cdots, \tag{2.6}$$

where $\mu_t = \mathbf{X}_t\beta_t$ is deterministic and the other components are stochastic [5].

Given observations $\{\mathbf{Z}_1, \cdots, \mathbf{Z}_T\}$ of a latent spatio-temporal process with the measurement error, the prediction of the latent spatio-temporal process $\{\mathbf{Y}_1, \cdots, \mathbf{Y}_T\}$ is a fundamental problem. Predicting the random variable $\mathbf{Y}_t$ on giving $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_t$ is called a filtering problem if $t = T$; a smoothing problem if $t < T$; and a prediction problem if $t > T$. Let $\mathcal{S}_t$ be the set of m_t spatial locations, $\mathbf{Y}_t$ be the vector of random variables $\{Y(\mathbf{s}; t)|\mathbf{s} \in \mathcal{S}_t\}$, and $\mathbf{Z}_t$ be vector of observations $\{Z(\mathbf{s}_1; t), \cdots, Z(\mathbf{s}_{n_t}; t)|\mathbf{s}_i \in \mathcal{S}_t, 1 \leq i \leq n_t\}$. Let $\eta_{t|\hat{t}} = E(\eta_t|\mathbf{Z}_{1:\hat{t}}), \xi_{t|\hat{t}} = E(\xi|\mathbf{Z}_{1:\hat{t}})$. Denote $\mathbf{P}_{t|\hat{t}} = Var(\eta_t|\mathbf{Z}_{1:\hat{t}})$ as the conditional covariance matrix of η_t and $\mathbf{R}_{t|\hat{t}} = Var(\xi|\mathbf{Z}_{1:\hat{t}})$ as the conditional covariance matrix of ξ.

Considering all of the component in Equation (2.6) are of the exponential family distribution, the result can be analytically solved by maximum likelihood estimation, which offers stable computation of valid estimators and makes efficient use of spatial and temporal dependence in the data.

The STRE filtering estimator is $\mathbf{Y}_{t|t}$:

$$\begin{aligned}
\mathbf{Y}_{t|t} &= \mathbf{O}_t\mathbf{X}_t\beta_t + \mathbf{O}_t\mathbf{S}_t^T\eta_{t|t} + \mathbf{O}_t\xi_{t|t}, \\
\eta_{t|t} &= \eta_{t|t-1} + \mathbf{P}_{t|t-1}\mathbf{S}_t\mathbf{O}_t^T(\mathbf{O}_t\mathbf{S}_t^T\mathbf{P}_{t|t-1}\mathbf{S}_t\mathbf{O}_t^T + \mathbf{D}_t)^{-1} \\
&\quad (\mathbf{Z}_t - \mathbf{O}_t\mathbf{X}_t\beta_t - \mathbf{O}_t\mathbf{S}_t^T\eta_{t|t-1}), \\
\xi_{t|t} &= \sigma_{\xi,t}^2\mathbf{V}_{\xi,t}\mathbf{O}_t^T(\mathbf{O}_t\mathbf{S}_t^T\mathbf{P}_{t|t-1}\mathbf{S}_t\mathbf{O}_t^T + \mathbf{D}_t)^{-1}(\mathbf{Z}_t - \mathbf{O}_t\mathbf{X}_t\beta_t - \mathbf{O}_t\mathbf{S}_t^T\eta_{t|t-1}), \\
\mathbf{P}_{t|t} &= \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{S}_t\mathbf{O}_t^T(\mathbf{O}_t\mathbf{S}_t^T\mathbf{P}_{t|t-1}\mathbf{S}_t\mathbf{O}_t^T + \mathbf{D}_t)^{-1}\mathbf{O}_t\mathbf{S}_t^T\mathbf{P}_{t|t-1}, \\
\mathbf{R}_{t|t} &= \sigma_{\xi,t}^2\mathbf{V}_{\xi,t} - \sigma_{\xi,t}^2\mathbf{V}_{\xi,t}\mathbf{O}_t^T(\mathbf{O}_t\mathbf{S}_t^T\mathbf{P}_{t|t-1}\mathbf{S}_t\mathbf{O}_t^T + \mathbf{D}_t)^{-1}\mathbf{O}_t\mathbf{V}_{\xi,t}\sigma_{\xi,t}^2,
\end{aligned} \tag{2.7}$$

where $\mathbf{D}_t = \sigma_{\xi,t}^2\mathbf{O}_t\mathbf{V}_{\xi,t}\mathbf{O}_t^T + \sigma_{\varepsilon,t}^2\mathbf{V}_{\varepsilon,t}$. The filtering process is usually conducted time stamp by time stamp, starting from time stamp 1. To conduct the filtering task for time t, we assume that the filtering task for time $t-1$ has been finished, hence $\eta_{t-1|t-1}$, $\eta_{t|t-1}$, and $\mathbf{P}_{t|t-1}$ are known.

The STRE smoothing estimator is $\mathbf{Y}_{t|T}, t < T$:

$$\begin{aligned}
\mathbf{Y}_{t|T} &= \mathbf{O}_t\mathbf{X}_t\beta_t + \mathbf{O}_t\mathbf{S}_t^T\eta_{t|T} + \mathbf{O}_t\xi_{t|T}, \\
\eta_{t|T} &= \eta_{t|t} + \mathbf{J}_t(\eta_{t+1|T} - \eta_{t+1|t}), \\
\xi_{t|T} &= \xi_{t|t} - \mathbf{M}_t(\eta_{t+1|T} - \eta_{t+1|t}), \\
\mathbf{P}_{t|T} &= \mathbf{P}_{t|t} + \mathbf{J}_t(\mathbf{P}_{t+1|T} - \mathbf{P}_{t+1|t})\mathbf{J}_t^T, \\
\mathbf{R}_{t|T} &= \mathbf{R}_{t|t} + \mathbf{M}_t(\mathbf{P}_{t+1|T} - \mathbf{P}_{t+1|t})\mathbf{M}_t^T, \\
\mathbf{J}_t &= \mathbf{P}_{t|t}\mathbf{H}_{t+1}^T\mathbf{P}_{t+1|t}^{-1}, \\
\mathbf{M}_t &= \sigma_{\xi,t}^2\mathbf{V}_{\xi,t}\mathbf{O}_t^T(\mathbf{O}_t\mathbf{S}_t^T\mathbf{P}_{t|t-1}\mathbf{S}_t\mathbf{O}_t^T + \mathbf{D}_t)^{-1}\mathbf{O}_t\mathbf{S}_t^T\mathbf{P}_{t|t-1}\mathbf{H}_{t+1}^T\mathbf{P}_{t+1|t}^{-1}.
\end{aligned} \tag{2.8}$$

where $\mathbf{D}_t$ is same as previous. This smoothing process is conducted based on filtering estimator. The present smoothed $\mathbf{Y}_{t|T}$ is related to $\mathbf{Y}_{t|t}$ and $\mathbf{Y}_{t+1|T}$.

After filtering $\mathbf{Y}_{t|t}$ and smoothing $\mathbf{Y}_{t|T}$ are estimated, they can be directly used to fill the missing area and get smoothed data. This STRE model fast process the large spatio-temporal data in linear time complexity and solve the analytical formulation. But the employed exponential family noise distribution might be sensitive to existing outliers or abnormal data.

2.2 The Forward-Backward Algorithm

The previous work of Spatio-Temporal Random Effect model utilized maximum likelihood estimation via an expectation-maximization algorithm to estimate the parameters and then substitute them into the optimal predictor. The likelihood calculation also can be solved using forward-backward algorithm. The forward-backward algorithm is an inference algorithm for sequential data, which calculates the marginal posterior of latent variables given the sequence of observations. The algorithm utilizes the principle of dynamic programming to efficiently calculate the values that are required to obtain the marginal posterior through two passes. The first pass goes forward in the time while the second goes backward in the time. This section illustrates the forward-backward algorithm for estimating η in the STRE model.

The spatio-temporal random effects model is described in Equations (2.5) and (2.6). Here, we re-formalize it as a state-space form

$$\begin{aligned} \mathbf{Z}_t &= \mathbf{O}_t\mathbf{X}_t\beta_t + \mathbf{O}_t\mathbf{S}_t^T\eta_t + \mathbf{O}_t\epsilon_t, \\ \eta_t &= \mathbf{H}_t\eta_{t-1} + \zeta_t, \end{aligned} \tag{2.9}$$

in which we define a new vector of variations ϵ_t, and each component ϵ_{tn} follows a Gaussian distribution with zero mean and the variance $\sigma_\xi^2 v_t'(\mathbf{s}) + \sigma_\varepsilon^2 v_t(\mathbf{s})$. Due to the independence structure of this state-space model, the joint distribution of latent states and observations is shown as:

$$p(\mathbf{Z}_1, \cdots, \mathbf{Z}_T, \eta_1, \cdots, \eta_T) = p(\eta_1)p(\mathbf{Z}_1|\eta_1)\prod_{t=2}^{T} p(\eta_t|\eta_{t-1})p(\mathbf{Z}_t|\eta_t), \tag{2.10}$$

where $p(\eta_1)$ is the prior of the first state [19].

The forward and backward messages can be defined as follows

$$\alpha_t(\eta_t) \equiv p(\mathbf{Z}_1, \cdots, \mathbf{Z}_t, \eta_t), \tag{2.11}$$

$$\beta_t(\eta_t) \equiv p(\mathbf{Z}_{t+1}, \cdots, \mathbf{Z}_T|\eta_t), \tag{2.12}$$

where message $\alpha_t(\eta_t)$ represents the joint probability of observations up to time t given η_t, and $\beta_t(\eta_t)$ represents the conditional probability of observations from time $t+1$ up to T given η_t. The recursive derivation of $\alpha_t(\eta_t)$ and $\beta_t(\eta_t)$ is presented in the following sections. The result is shown as

$$\alpha_t(\eta_t) = p(\mathbf{Z}_t|\eta_t) \int p(\eta_t|\eta_{t-1})\alpha_{t-1}(\eta_{t-1})d\eta_{t-1}, \tag{2.13}$$

$$\beta_t(\eta_t) = \int p(\mathbf{Z}_{t+1}|\eta_{t+1})p(\eta_{t+1}|\eta_t)\beta_{t+1}(\eta_{t+1})d\eta_{t+1}. \tag{2.14}$$

As indicated in Equation (2.10), in order to calculate the density $p(\eta_t|\mathbf{Z}_{1:T})$, we need to integrate out all the other latent states. According to the definition of $\alpha_t(\eta_t)$, $\beta_t(\eta_t)$, we can derive the one-slice and two-slice marginal distributions of latent variables as follows [19]:

$$p(\eta_t|\mathbf{Z}_{1:T}) = \frac{1}{p(\mathbf{Z}_{1:T})}\alpha_t(\eta_t)\beta_t(\eta_t), \tag{2.15}$$

$$p(\eta_{t-1},\eta_t|\mathbf{Z}_{1:T}) = \frac{1}{p(\mathbf{Z}_{1:T})}\alpha_{t-1}(\eta_{t-1})p(\eta_t|\eta_{t-1})p(\mathbf{Z}_t|\eta_t)\beta_t(\eta_t). \tag{2.16}$$

2.2.1 Forward Recursion

This and the following section will show the derivation of the recursion relations of $\alpha_t(\eta_t)$ and $\beta_t(\eta_t)$, which are allowed to evaluated them efficiently.

In Figure 2.1, we use the discrete hidden states to illustrate the forward recursion, which can be generic to continuous hidden states. At time $t-1$, $\alpha_{t-1}(\eta_{t-1})$ collects messages of each hidden states and its weights given by the transition probability, and then multiplies them by data contribution $p(\mathbf{Z}_t|\eta_t)$, sends this to $\alpha_t(\eta_t)$.

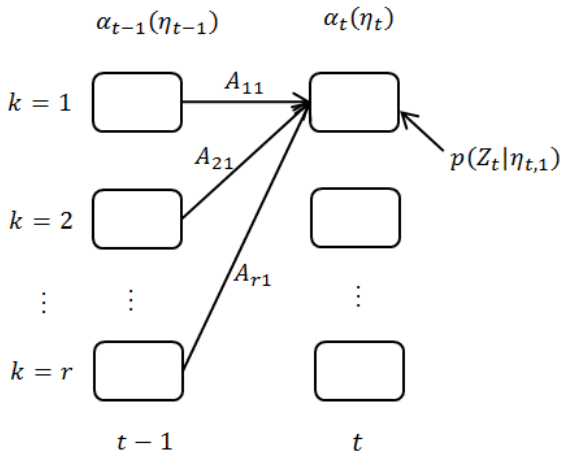


Figure 2.1: Illustration of forward recursion

From the definition of $\alpha_t(\eta_t)$, we try to express it in terms of $\alpha_{t-1}(\eta_{t-1})$ as Equation (2.11):

$$\begin{aligned}
\alpha_t(\eta_t) &= p(\mathbf{Z}_1, \cdots, \mathbf{Z}_t, \eta_t) \\
&= p(\mathbf{Z}_1, \cdots, \mathbf{Z}_t|\eta_t)p(\eta_t) \\
&= p(\mathbf{Z}_t|\eta_t)p(\mathbf{Z}_1, \cdots, \mathbf{Z}_{t-1}|\eta_t)p(\eta_t) \\
&= p(\mathbf{Z}_t|\eta_t)p(\mathbf{Z}_1, \cdots, \mathbf{Z}_{t-1}, \eta_t) \\
&= p(\mathbf{Z}_t|\eta_t)\int p(\mathbf{Z}_1, \cdots, \mathbf{Z}_{t-1}, \eta_{t-1}, \eta_t)d\eta_{t-1} \\
&= p(\mathbf{Z}_t|\eta_t)\int p(\mathbf{Z}_1, \cdots, \mathbf{Z}_{t-1}, \eta_t|\eta_{t-1})p(\eta_{t-1})d\eta_{t-1} \\
&= p(\mathbf{Z}_t|\eta_t)\int p(\mathbf{Z}_1, \cdots, \mathbf{Z}_{t-1}|\eta_{t-1})p(\eta_t|\eta_{t-1})p(\eta_{t-1})d\eta_{t-1} \\
&= p(\mathbf{Z}_t|\eta_t)\int p(\mathbf{Z}_1, \cdots, \mathbf{Z}_{t-1}, \eta_{t-1})p(\eta_t|\eta_{t-1})d\eta_{t-1} \\
&= p(\mathbf{Z}_t|\eta_t)\int p(\eta_t|\eta_{t-1})\alpha_{t-1}(\eta_{t-1})d\eta_{t-1}
\end{aligned}$$

where $\alpha_{t-1}(\eta_{t-1}) = p(\mathbf{Z}_1, \ldots, \mathbf{Z}_{t-1}, \eta_{t-1})$ from the definition. So we get the recursive Equation (2.13).

After getting the recursion relations of $\alpha_t(\eta_t)$, it is assumed that $\alpha_t(\eta_t)$ follows gaussian distribution: $\alpha_t(\eta_t) \sim \mathcal{N}(\eta_{t|t}, \mathbf{P}_{t|t})$. The recursive forward equation (2.13) becomes:

$$\begin{aligned}
\alpha_t(\eta_t) &= \mathcal{N}(\mathbf{Z}_t; \mathbf{O}_t\mathbf{X}_t\beta_t + \mathbf{O}_t\mathbf{S}_t^T\eta_t, \mathbf{D}_t) \\
&\quad \int \mathcal{N}(\eta_t; \mathbf{H}_t\eta_{t-1}, \mathbf{U}_t)\mathcal{N}(\eta_{t-1}; \eta_{t-1|t-1}, \mathbf{P}_{t-1|t-1})d\eta_{t-1}, \qquad (2.17)
\end{aligned}$$

From this equation, we can apply the principle of product and conditional distribution of gaussian distribution, and the mean and variance of $\alpha_t(\eta_t)$ are given as:

$$\begin{aligned} \eta_{t|t} &= \mathbf{H}_t\eta_{t-1|t-1} + \mathbf{K}_t(\mathbf{Z}_t - \mathbf{O}_t\mathbf{X}_t\beta_t - \mathbf{O}_t\mathbf{S}_t^T\mathbf{H}_t\eta_{t-1|t-1}), \\ \mathbf{P}_{t|t} &= (\mathbf{I} - \mathbf{K}_t\mathbf{O}_t\mathbf{S}_t^T)\mathbf{P}_{t|t-1}, \end{aligned} \tag{2.18}$$

where

$$\begin{aligned} \mathbf{K}_t &= \mathbf{P}_{t|t-1}\mathbf{S}_t\mathbf{O}_t^T(\mathbf{O}_t\mathbf{S}_t^T\mathbf{P}_{t|t-1}\mathbf{S}_t\mathbf{O}_t^T + \mathbf{D}_t)^{-1}, \\ \mathbf{P}_{t|t-1} &= \mathbf{H}_t\mathbf{P}_{t-1|t-1}\mathbf{H}_t^T + \mathbf{U}_t. \end{aligned}$$

The message $\alpha_1(\eta_1) = p(\mathbf{Z}_1, \eta_1) = p(\eta_1)p(\mathbf{Z}_1|\eta_1)$, and the recursions start with

$$\begin{aligned} \eta_{1|1} &= \mathbf{H}_1\eta_{0|0} + \mathbf{K}_1(\mathbf{Z}_1 - \mathbf{O}_1\mathbf{X}_1\beta_1 - \mathbf{O}_1\mathbf{S}_1^T\mathbf{H}_1\eta_{0|0}), \\ \mathbf{P}_{1|1} &= (\mathbf{I} - \mathbf{K}_1\mathbf{O}_1\mathbf{S}_1^T)\mathbf{P}_{1|0}, \end{aligned}$$

where

$$\begin{aligned} \mathbf{K}_1 &= \mathbf{P}_{1|0}\mathbf{S}_1\mathbf{O}_1^T(\mathbf{O}_1\mathbf{S}_1^T\mathbf{P}_{1|0}\mathbf{S}_1\mathbf{O}_1^T + \mathbf{D}_1)^{-1}, \\ \mathbf{P}_{1|0} &= \mathbf{H}_1\mathbf{P}_{0|0}\mathbf{H}_1^T + \mathbf{U}_1, \end{aligned}$$

where $\eta_{0|0}$ and $\mathbf{P}_{0|0}$ are the initial values.

By given the initial states of η and $\mathbf{P}$, all of forward messages $\alpha_t(\eta_t)$ can be easily calculated by using the above forward message passing steps.

2.2.2 Backward Recursion

In the Figure 2.2, we illustrate the backward recursion after forward recursion. At time $t+1$, $\beta_{t+1}(\eta_{t+1})$ collects message of each hidden states multiplies by data contribution $p(\mathbf{Z}_{t+1}|\eta_{t+1})$, and then obtains their weights given by the transition probability, sends back to $\beta_t(\eta_t)$.

From the definition of $\beta_{t+1}(\eta_{t+1})$, we try to express it in terms of $\beta_t(\eta_t)$ as Equation (2.12):

$$\begin{aligned} \beta_t(\eta_t) &= p(\mathbf{Z}_{t+1}, \cdots, \mathbf{Z}_T|\eta_t) \\ &= \int p(\mathbf{Z}_{t+1}, \cdots, \mathbf{Z}_T, \eta_{t+1}|\eta_t)d\eta_{t+1} \\ &= \int p(\mathbf{Z}_{t+1}, \cdots, \mathbf{Z}_T|\eta_t, \eta_{t+1})p(\eta_{t+1}|\eta_t)d\eta_{t+1} \\ &= \int p(\mathbf{Z}_{t+1}, \cdots, \mathbf{Z}_T|\eta_{t+1})p(\eta_{t+1}|\eta_t)d\eta_{t+1} \\ &= \int p(\mathbf{Z}_{t+2}, \cdots, \mathbf{Z}_T|\eta_{t+1})p(\mathbf{Z}_{t+1}|\eta_{t+1})p(\eta_{t+1}|\eta_t)d\eta_{t+1} \\ &= \int p(\mathbf{Z}_{t+1}|\eta_{t+1})p(\eta_{t+1}|\eta_t)\beta_{t+1}(\eta_{t+1})d\eta_{t+1} \end{aligned}$$

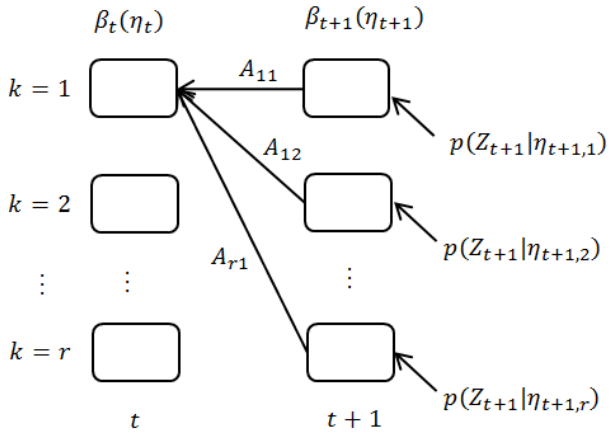


Figure 2.2: Illustration of backward recursion

where $\beta_{t+1}(\eta_{t+1}) = p(\mathbf{Z}_{t+2}, \cdots, \mathbf{Z}_T|\eta_{t+1})$ from the definition. So we get the recursive equation (2.14).

The derivation of the sequential backward recursion after the forward recursion through time T is by: Given that $\alpha_t(\eta_t)\beta_t(\eta_t) \sim \mathcal{N}(\eta_t; \eta_{t|T}, \mathbf{P}_{t|T})$, we utilize Equations (2.14), (2.18) to acquire the following formula:

$$\begin{aligned} \alpha_t(\eta_t)\beta_t(\eta_t) &= \mathcal{N}(\eta_t; \eta_{t|t}, \mathbf{P}_{t|t}) \int d\eta_{t+1} \mathcal{N}(\eta_{t+1}; \mathbf{H}_{t+1}\eta_t, \mathbf{U}_{t+1}) \\ &\quad \mathcal{N}(\mathbf{Z}_{t+1}; \mathbf{O}_{t+1}\mu_{t+1} + \mathbf{O}_{t+1}\mathbf{S}_{t+1}^T\eta_{t+1}, \mathbf{D}_{t+1}) \mathcal{N}(\eta_{t+1}; \eta_{t+1|T}, \mathbf{P}_{t+1|T}), \end{aligned} \tag{2.19}$$

Similar to the forward recursion process, the mean and variance of $\alpha_t(\eta_t)\beta_t(\eta_t)$ can be derived by using the rules of product and conditional distribution of gaussian distribution.

$$\begin{aligned} \eta_{t|T} &= \eta_{t|t} + \mathbf{J}_t(\eta_{t+1|T} - \mathbf{H}_{t+1}\eta_{t|t}), \\ \mathbf{P}_{t|T} &= \mathbf{P}_{t|t} + \mathbf{J}_t(\mathbf{P}_{t+1|T} - \mathbf{P}_{t+1|t})\mathbf{J}_t^T, \end{aligned} \tag{2.20}$$

where

$$\mathbf{J}_t = \mathbf{P}_{t|t}\mathbf{H}_{t+1}^T\mathbf{P}_{t+1|t}^{-1}.$$

Since $\beta_T(\eta_T) = 1$, the backward recursion starts with $\alpha_T(\eta_T)$.

After we acquire the value of $\alpha_t(\eta_t)$ and $\beta_t(\eta_t)$, and the same process for the $\xi_{t|t}$ and $\xi_{t|T}$. The STRE filtering and smoothing results can easily calculated by Equations (2.18), (2.20).

2.3 Expectation Propagation

Expectation Propagation (EP) [20] is an efficient approximate inference framework that has shown better predictive performance than traditional inference approaches, such as variational approximation and Laplace approximation [23]. The main ideas of Expectation Propagation is finding some Gaussian distribution to moment match the original distribution. Given observed data $\mathcal{D}$ and hidden variables (including parameters) θ, for many probabilistic models, the posterior distribution of θ given $\mathcal{D}$ comprises a product of factors with the form

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D},\theta)}{p(\mathcal{D})} = \frac{1}{p(\mathcal{D})}\prod_i f_i(\theta). \tag{2.21}$$

and the model evidence is given by

$$p(\mathcal{D}) = \int \prod_i f_i(\theta) d\theta.$$

In an independent, identically distributed data, the factor $f_i(\theta) = p(\mathbf{x}_i|\theta)$ for each data point $\mathbf{x}_i$, along with a factor $f_0(\theta) = p(\theta)$ corresponding to the prior.

In this scenario, the most interesting parts are to evaluate the posterior for the purpose of making predictions, as well as the evidence $p(\mathcal{D})$ for the purpose of model comparison.

Expectation Propagation aims to approximate $p(\theta|\mathcal{D})$ by a product of factors

$$q(\theta) = \frac{1}{p(\mathcal{D})}\prod_i \tilde{f}_i(\theta), \tag{2.22}$$

in which each factor $\tilde{f}_i(\theta)$ corresponds to the one of the factors $f_i(\theta)$ in Equation (2.21). The factors $\tilde{f}_i(\theta)$ are usually constrained to parametric forms (e.g., exponential family) in order to make the inference algorithm practical.

In Figure 2.3, it shows the approximation scenario of expectation propagation. The blue curves show the exact distribution and the green ones present the approximated distribution. If taking the exact $f_i(x)$ and approximated $\tilde{f}_i(x)$ to multiply $q^{\setminus i}(x)$, the results show that two different multiplications are highly close. This motivates us to iteratively remove one factor from the approximated Gaussian distribution, then approximate the new distribution of remaining part and newly added real distribution of f_i. After approximating the whole distribution, we can use the new approximated Gaussian distribution to divide the remaining part to calculate the approximated $\tilde{f}_i$.

Basically, Expectation Propagation conducts iterative refinement of the approximate posterior $q(\theta|\mathcal{D})$ by adding additional message passing through the factors. For each iteration, EP first replaces one of the approximate factors $\tilde{f}_i(\theta)$ with the true factor $f_i(\theta)$, the result denoted as $q^{\setminus i}(\theta)f_i(\theta)$ in which $q^{\setminus i}(\theta)$ denotes the proposed $q(\theta)$ with the ith factors $\tilde{f}_i(\theta)$ removed.

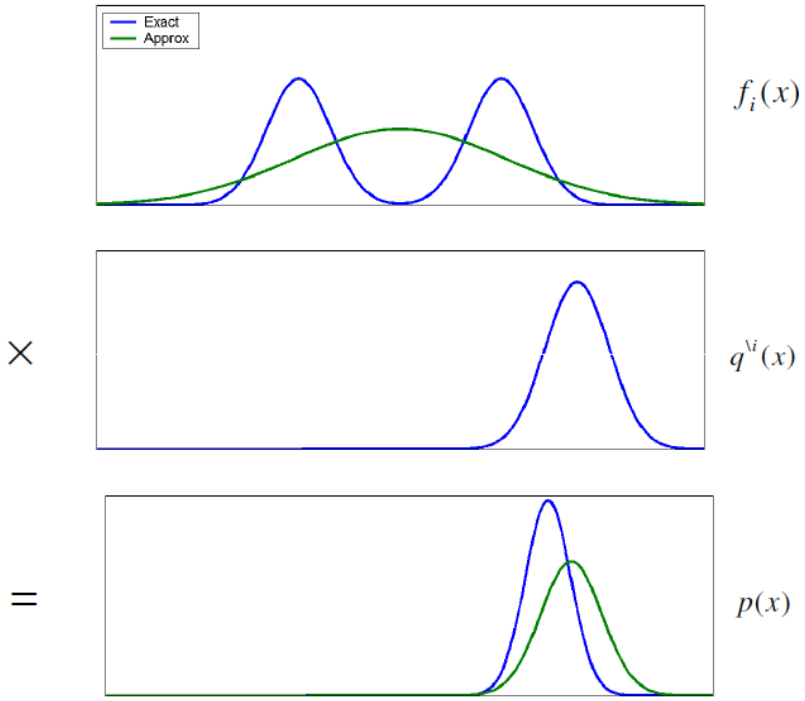


Figure 2.3: Intuition of Expectation Propagation

It then refines the new posterior by moments matching between $q^{new}(\theta)$ and $q^{\setminus i}(\theta)f_i(\theta)$. The new factor $\tilde{f}_i(\theta)$ is conducted by making sure the product

$$q^{new}(\theta) \propto \tilde{f}_i(\theta) \prod_{j \neq i} \tilde{f}_j(\theta)$$

is as close as possible to

$$f_i(\theta) \prod_{j \neq i} \tilde{f}_j(\theta)$$

After that, the new factor $\tilde{f}_i(\theta)$ is updated as

$$\tilde{f}_i(\theta) \propto \frac{q^{new}(\theta)}{q^{\setminus i}(\theta)}. \tag{2.23}$$

Expectation Propagation continues the refinement iterations until all factors $\tilde{f}_i(\theta)$ converge. Note that, Expectation Propagation convergence has not been theoretically justified, but in practice the convergence is often achieved as occurred in our problem.

The complete procedure of Expectation Propagation is presented as follows.

Algorithm 1: Expectation Propagation

Given a joint distribution over observed data $\mathcal{D}$ and stochastic variables θ in the form of a product of factors

$$p(\mathcal{D}, \theta) = \prod_i f_i(\theta)$$

The objective is to approximate the posterior distribution $p(\theta|\mathcal{D})$ by

$$q(\theta) \propto \prod_i \tilde{f}_i(\theta)$$

1. Initialize all of the approximating factors $\tilde{f}_i(\theta)$.
2. Initialize all posterior approximation by setting

$$q(\theta) \propto \prod_i \tilde{f}_i(\theta).$$

3. Until convergence:
   (a) Choose a factor $\tilde{f}_j(\theta)$ to refine.
   (b) Remove $\tilde{f}_j(\theta)$ from the posterior by division

$$q^{\setminus j}(\theta) \propto \frac{q(\theta)}{\tilde{f}_j(\theta)}.$$

   (c) Evaluate the new posterior by setting the sufficient statistics of $q^{new}(\theta)$ equal to those of $q^{\setminus j} f_j(\theta)$.
   (d) Evaluate and store the new factor

$$\tilde{f}_j(\theta) \propto \frac{q^{new}(\theta)}{q^{\setminus j}(\theta)}$$

Expectation Propagation can not guarantee the iterations to be converged. However, if the iterations of approximation of $q(\theta)$ to exponential family do converge, the result will be a stationary status.

Chapter 3

Student-t Robust Spatio-Temporal Prediction Model

In this chapter, we first introduce the new Student-t Robust Spatio-Temporal Prediction model, and then present solutions to estimate the posterior distributions $p(\mathbf{Y}_t|\mathbf{Z}_{1:t})$ and $p(\mathbf{Y}_t|\mathbf{Z}_{1:T})$ by separately estimating the η and ξ.

3.1 Student-t RSTP Model

This section presents the Student-t Robust Spatio-Temporal Prediction (St-RSTP) model, which considers Student's t-distribution to model the measurement error, instead of the traditional Gaussian distribution. As shown in Figure 1, Student's t-distribution has a heavier tail than Gaussian distribution. The tail heaviness is controlled by setting the degrees of freedom (ν), and when the degree of freedom approaches infinity, Student's t-distribution becomes equivalent to Gaussian distribution. Student's t-distribution has been used in a number of statistical models and has been shown effective for a variety of robust processes [7, 23]. We use the same symbols as in Chapter 2.1. The St-RSTP model can be formalized as

$$\mathbf{Z}_t = \mathbf{O}_t\mathbf{Y}_t + \varepsilon_t, \tag{3.1}$$

$$\mathbf{Y}_t = \mathbf{X}_t\beta_t + \mathbf{S}_t^T\eta_t + \xi_t, \tag{3.2}$$

$$\eta_t = \mathbf{H_t}\eta_{t-1} + \zeta_t, \tag{3.3}$$

where $\mathbf{O}_t$ is the incidence matrix, $\mathbf{Y}_t$ is the vector of latent variables to be predicted, and $\mathbf{X}_t$ is the matrix of co-variates, β_t is the coefficients, which is generally known. $\mathbf{S}_t$ is a vector of basis functions, η_t captures the small scale variation and is modeled as the vector-autoregressive process (see Equation (2.5)), ξ_t captures a micro-scale variation, and each ξ_{tn}, $n = 1 \cdots, m_t$, is modeled by a white noise Gaussian process with mean zero and variance

$\sigma_\xi^2 v_t'(\mathbf{s})$. As a key difference from the STRE model, the measurement error ε_{tn} now follows a Student's t-distribution $\textit{Student-t}(0, \nu, \sigma)$ with the probability density function as

$$p(\varepsilon_{tn}) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)}(\frac{1}{\pi\nu\sigma})^{\frac{1}{2}}(1+\frac{\varepsilon_{tn}^2}{\nu\sigma})^{-\frac{\nu}{2}-\frac{1}{2}}, \tag{3.4}$$

where ν is the degrees of freedom and σ is the scale parameter.

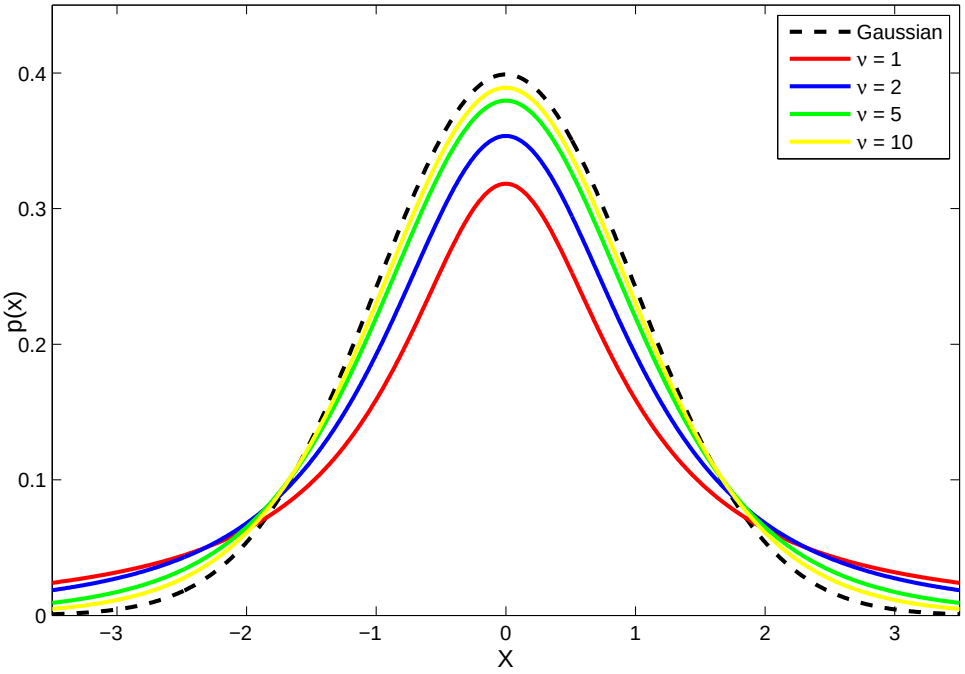


Figure 3.1: Student's t vs. Gaussian Distribution

Given observations $\{\mathbf{Z}_1, ..., \mathbf{Z}_T\}$, the predictive process is to estimate the latent variables $\{\mathbf{Y}_1, ..., \mathbf{Y}_T\}$ at sampled and unsampled locations. The prediction of $\mathbf{Y}$ variables at unsampled locations is realized by using the incidence matrix $\mathbf{O}_t$ in the St-RSTP model, where $\mathbf{O}_t \in \mathcal{R}^{n_t \times m_t}$, n_t refers to the number of observations at sampled locations, and $m_t - n_t$ refers to the number of unsampled locations that are of interest for prediction. As discussed in Chapter 2.1, the prediction at previous time, the current time, and the future time, are named smoothing, filtering, and forecasting, respectively. This thesis focuses on smoothing and filtering processes. The objective of this work is to estimate the expectation and variance-covariance of the posterior distributions $p(\mathbf{Y}_t|\mathbf{Z}_{1:T}), t = 1, \cdots, T$, denoted as $\mathbf{Y}_{t|T}$ and $\mathbf{\Sigma}_{t|T}$, respectively. $\mathbf{Y}_{t|T}$ will be used as the prediction values, and $\mathbf{\Sigma}_{t|T}$ will be used to estimate the confidence interval. By using the similar strategy as in the regular STRE model (Chapter 2.1), we firstly estimate the posterior distributions of the two components

$$p(\eta_t|\mathbf{Z}_{1:T}), p(\xi_t|\mathbf{Z}_{1:T}), \tag{3.5}$$

where we denote $\eta_{t|T} \equiv E[p(\eta_t|\mathbf{Z}_{1:T})]$, $\mathbf{P}_{t|T} \equiv Var[p(\eta_t|\mathbf{Z}_{1:T})]$, $\xi_{t|T} \equiv E[p(\xi_t|\mathbf{Z}_{1:T})]$, and $\mathbf{R}_{t|T} \equiv Var[p(\xi_t|\mathbf{Z}_{1:T})]$. Then, it follows that

$$\begin{aligned} \mathbf{Y}_{t|T} &= \eta_{t|T} + \xi_{t|T}, \\ \mathbf{\Sigma}_{t|T} &= \mathbf{P}_{t|T} + \mathbf{R}_{t|T}. \end{aligned} \tag{3.6}$$

Figure 3.2 gives the graph representation of the St-RSTP model. The following sections discuss the estimations of $\eta_{t|T}$ and $\mathbf{P}_{t|T}$, and the estimations of $\xi_{t|T}$ and $\mathbf{R}_{t|T}$.

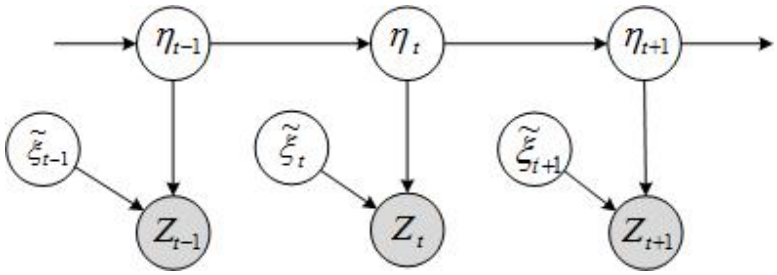


Figure 3.2: St-RSTP Graphic Model

3.2 Approximating $p(\eta_t|\mathbf{Z}_{1:t})$ and $p(\eta_t|\mathbf{Z}_{1:T})$

In order to reduce the dimensionality of the inference process, we estimate the posterior distributions of η_t and ξ_t separately, by using a similar strategy as in the original STRE paper [5]. Recall that a standard STRE data process has the form (Equation (2.6))

$$\mathbf{Z}_t = \mathbf{O}_t\mathbf{X}_t\beta_t + \mathbf{O}_t\mathbf{S}_t^T\eta_t + \mathbf{O}_t\xi_t + \varepsilon_t.$$

Each component ξ_{tn} captures a micro-scale variation and is modeled by a white noise Gaussian process with mean zero and variance $var(\xi(\mathbf{s};t)) = \sigma_\xi^2 v_t'(\mathbf{s})$. ε_{tn} is a spatial white-noise Gaussian process with mean zero and variance $var(\varepsilon(\mathbf{s};t)) = \sigma_\varepsilon^2 v_t(\mathbf{s})$. The incidence matrix $\mathbf{O}_t$ is used to handle missing values that are related to locations where no observations are available. The product $\mathbf{O}_t\xi_t$ aims to remove the components in ξ_t that do not have related observations. By using this feature, we define a new vector of variations ϵ_t, where each component ϵ_{tn} follows a Gaussian distribution with zero mean and the variance $\sigma_\xi^2 v_t'(\mathbf{s}) + \sigma_\varepsilon^2 v_t(\mathbf{s})$. The sum $\sigma_\xi^2 v_t'(\mathbf{s}) + \sigma_\varepsilon^2 v_t(\mathbf{s})$ is usually called the "nugget effect" in the geostatistical literature [5]. Following that, the STRE model can be reformulated as the form

$$\mathbf{Z}_t = \mathbf{O}_t\mathbf{X}_t\beta_t + \mathbf{O}_t\mathbf{S}_t^T\eta_t + \mathbf{O}_t\epsilon_t. \tag{3.7}$$

Theorem 1. *The posterior distributions $p(\eta_t|\mathbf{Z}_{1:t})$ and $p(\eta_t|\mathbf{Z}_{1:T})$ estimated based on the form (3.7) are equivalent to those estimated based on the regular form (2.6)*

Proof. First, the posterior distributions $p(\eta_t|\mathbf{Z}_{1:t})$ and $p(\eta_t|\mathbf{Z}_{1:T})$ are estimated based on the formulation as shown in Equation (3.7), which implies that

$$\begin{aligned}\mathbf{Z}_t &= \mathbf{O}_t\mathbf{X}_t\beta_t + \mathbf{O}_t\mathbf{S}_t^T\eta_t + \mathbf{O}_t\epsilon_t,\\ \eta_t &= \mathbf{H}_t\eta_{t-1} + \zeta_t.\end{aligned}$$

Given the initial values $\eta_{0|0}$ and $\mathbf{P}_{0|0}$, it can be readily derived that [25]

$$\begin{aligned}\eta_{t|t} &= \eta_{t|t-1} + \mathbf{P}_{t|t-1}\mathbf{S}_t\mathbf{O}_t^T(\mathbf{O}_t\mathbf{S}_t^T\mathbf{P}_{t|t-1}\mathbf{S}_t\mathbf{O}_t^T + \mathbf{D}_t)^{-1},\\ \mathbf{P}_{t|t} &= \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{S}_t\mathbf{O}_t^T(\mathbf{O}_t\mathbf{S}_t^T\mathbf{P}_{t|t-1}\mathbf{S}_t\mathbf{O}_t^T + \mathbf{D}_t)^{-1}\mathbf{O}_t\mathbf{S}_t^T\mathbf{P}_{t|t-1}.\end{aligned} \tag{3.8}$$

Second, $p(\eta_t|\mathbf{Z}_{1:t})$ and $p(\eta_t|\mathbf{Z}_{1:T})$ are estimated based on the regular STRE formulation as shown in Equation (2.6), which implies that

$$\begin{aligned}\mathbf{Z}_t &= \mathbf{O}_t\mathbf{X}_t\beta_t + \mathbf{O}_t\mathbf{S}_t^T\eta_t + \mathbf{O}_t\xi_t + \varepsilon_t,\\ \eta_t &= \mathbf{H}_t\eta_{t-1} + \zeta_t.\end{aligned}$$

According to the definition of $\eta_{t|t-1}$, it follows that

$$\begin{aligned}\eta_{t|t-1} &= E(\eta_t|\mathbf{Z}_{1:t-1}) = E[(\mathbf{H}_t\eta_{t-1} + \zeta_t)|\mathbf{Z}_{1:t-1}] = \mathbf{H}_t\eta_{t-1|t-1},\\ \mathbf{P}_{t|t-1} &= E\{(\eta_t - \eta_{t|t-1})(\eta_t - \eta_{t|t-1})^T\}\\ &= E\left\{[\mathbf{H}_t(\eta_{t-1} - \eta_{t-1|t-1}) + \zeta_t][\mathbf{H}_t(\eta_{t-1} - \eta_{t-1|t-1}) + \zeta_t]^T\right\}\\ &= \mathbf{H}_t\mathbf{P}_{t-1|t-1}\mathbf{H}_t^T + \mathbf{U}_t.\end{aligned} \tag{3.9}$$

To derive Equation (2.7), we first estimate the prediction error δ_t and its variance as

$$\begin{aligned}\delta_t &= \mathbf{Z}_t - E(\mathbf{Z}_t|\mathbf{Z}_{1:t-1}) = \mathbf{Z}_t - (\mathbf{O}_t\mathbf{X}_t\beta_t + \mathbf{O}_t\mathbf{S}_t^T\eta_{t|t-1})\\ &= \mathbf{O}_t\mathbf{S}_t^T(\eta_t - \eta_{t|t-1}) + \mathbf{O}_t\xi_t + \varepsilon_t,\\ \mathbf{\Delta}_t &= Var(\delta_t) = \mathbf{O}_t\mathbf{S}_t^T\mathbf{P}_{t|t-1}\mathbf{S}_t\mathbf{O}_t^T + \mathbf{D}_t,\\ \mathbf{D}_t &= \sigma_{\xi,t}^2\mathbf{O}_t\mathbf{V}_{\xi,t}\mathbf{O}_t^T + \sigma_{\varepsilon,t}^2\mathbf{V}_{\varepsilon,t}.\end{aligned} \tag{3.10}$$

The fact $E(\delta_t\mathbf{Z}_s^T) = 0$, for $s < t$, implies that the prediction errors are independent of past observations. The conditional covariance of η_t and δ_t given $\mathbf{Z}_{1:t-1}$ has the form

$$\begin{aligned}cov(\eta_t, \delta_t|\mathbf{Z}_{1:t-1}) &= cov(\eta_t, \mathbf{Z}_t - \mathbf{O}_t\mathbf{X}_t\beta_t - \mathbf{O}_t\mathbf{S}_t^T\eta_{t|t-1}|\mathbf{Z}_{1:t-1})\\ &= cov[\eta_t - \eta_{t|t-1}, \mathbf{O}_t\mathbf{S}_t^T(\eta_t - \eta_{t|t-1}) + \mathbf{O}_t\xi_t + \varepsilon_t]\\ &= \mathbf{P}_{t|t-1}\mathbf{S}_t\mathbf{O}_t^T.\end{aligned}$$

Using the above results, the joint conditional distribution of η_t and δ_t given $\mathbf{Z}_{1:t-1}$ follows a Gaussian distribution with the form

$$\left(\begin{array}{c} \eta_t \\ \delta_t \end{array} \right) \Bigg| \mathbf{Z}_{1:t-1} \sim \mathcal{N} \left(\left[\begin{array}{c} \eta_{t|t-1} \\ \mathbf{0} \end{array} \right], \left[\begin{array}{cc} \mathbf{P}_{t|t-1} & \mathbf{P}_{t|t-1}\mathbf{S}_t\mathbf{O}_t^T \\ \mathbf{O}_t\mathbf{S}_t^T\mathbf{P}_{t|t-1} & \mathbf{\Delta}_t \end{array} \right] \right).$$

Thus, using the (B.9) in appendix B of [25] and plugging into Equation (3.10), we can write

$$\begin{aligned} \eta_{t|t} &= E(\eta_t|\mathbf{Z}_{1:t-1}, \mathbf{Z}_t) = E(\eta_t|\mathbf{Z}_{1:t-1}, \delta_t) \\ &= \eta_{t|t-1} + \mathbf{P}_{t|t-1}\mathbf{S}_t\mathbf{O}_t^T(\mathbf{O}_t\mathbf{S}_t^T\mathbf{P}_{t|t-1}\mathbf{S}_t\mathbf{O}_t^T + \mathbf{D}_t)^{-1}(\mathbf{Z}_t - \mathbf{O}_t\mathbf{X}_t\beta_t - \mathbf{O}_t\mathbf{S}_t^T\eta_{t|t-1}), \\ \mathbf{P}_{t|t} &= cov(\eta_t|\mathbf{Z}_{1:t-1}, \delta_t) \\ &= \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{S}_t\mathbf{O}_t^T\mathbf{\Delta}_t^{-1}\mathbf{O}_t\mathbf{S}_t^T\mathbf{P}_{t|t-1} \\ &= \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{S}_t\mathbf{O}_t^T(\mathbf{O}_t\mathbf{S}_t^T\mathbf{P}_{t|t-1}\mathbf{S}_t\mathbf{O}_t^T + \mathbf{D}_t)^{-1}\mathbf{O}_t\mathbf{S}_t^T\mathbf{P}_{t|t-1}. \end{aligned} \tag{3.11}$$

Combining Equations (3.8) and (3.11), the posterior distributions $p(\eta_t|\mathbf{Z}_{1:t})$ estimated based on the form (3.7) are equivalent to those estimated based on the regular form (2.6). The proof of $p(\eta_t|\mathbf{Z}_{1:T})$ can be derived base on the result of $p(\eta_t|\mathbf{Z}_{1:t})$, with the similar procedures as used in the estimation of $p(\eta_t|\mathbf{Z}_{1:t})$. □

By using this good property of the STRE model, we can reformulate our St-STRP model as the similar form as in (3.7), by modeling the each component of ϵ_t as a Student's t-distribution (See Equation (3.4)). The posterior distributions $p(\eta_t|\mathbf{Z}_{1:t})$ and $p(\eta_t|\mathbf{Z}_{1:T})$ can be estimated accordingly.

In order to apply EP to the estimation problem, we first present the factor graph representation in the framework of dynamic Bayesian networks as shown in Figure 3.3. Factor graphs are bipartite graph that expresses the structure of factorization. The factor graph contains two distinct kinds of nodes: the variable nodes for each variable η_t; the factor nodes for factor functions Ω_t [4]. All links go between nodes of opposite type. The joint distribution of latent variables and observations is the product of all factor functions, as shown in Equation (3.13).

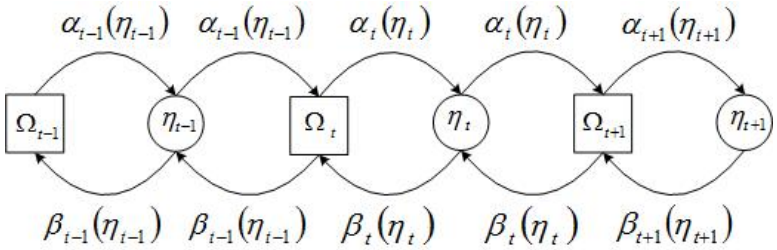


Figure 3.3: Factor Graph Presentation of St-STRE

From Figure 3.3, the joint distribution of latent variables and observations, forward and backward message passing components $\alpha(\cdot)$ and $\beta(\cdot)$ can be derived from Chapter 2.3, as showed below:

$$\begin{aligned} p(\eta_{1:T}, \mathbf{Z}_{1:T}) &= p(\eta_1)p(\mathbf{Z}_1|\eta_1)\prod_{t=2}^{T} p(\eta_t|\eta_{t-1})p(\mathbf{Z}_t|\eta_t), \\ \alpha_t(\eta_t) &= p(\mathbf{Z}_t|\eta_t)\int p(\eta_t|\eta_{t-1})\alpha_{t-1}(\eta_{t-1})d\eta_{t-1}, \\ \beta_{t-1}(\eta_{t-1}) &= \int p(\eta_t|\eta_{t-1})p(\mathbf{Z}_t|\eta_t)\beta_t(\eta_t)d\eta_t. \end{aligned} \tag{3.12}$$

The posterior distribution of latent variable can be re-formalized as the production of factor functions:

$$\begin{aligned} p(\eta_{1:T}|\mathbf{Z}_{1:T}) &= \frac{1}{p(\mathbf{Z}_{1:T})}p(\eta_1)p(\mathbf{Z}_1|\eta_1)\prod_{t=2}^{T} p(\eta_t|\eta_{t-1})p(\mathbf{Z}_t|\eta_t) \\ &\propto \prod_t \Omega_t(\eta_{t-1}, \eta_t), \end{aligned} \tag{3.13}$$

where each factor function is represented as

$$\Omega_t(\eta_{t-1}, \eta_t) := p(\eta_t|\eta_{t-1})p(\mathbf{Z}_t|\eta_t),$$

and $\Omega_t(\eta_0, \eta_1) := p(\eta_1)p(\mathbf{Z}_1|\eta_1)$, when $t = 1$.

Recall that $p(\mathbf{Z}_t|\eta_t)$ follows a Student's t-distribution, the estimation of Equation (3.12) is intractable. It can be further approximated as the following factorized form

$$q(\eta) = \prod_t q_t(\eta_{t-1}, \eta_t) \propto \prod_t \hat{\Omega}(\eta_{t-1}, \eta_t), \tag{3.14}$$

where $\hat{\Omega}(\eta_{t-1}, \eta_t)$ is an approximation of the factor $\Omega_t(\eta_{t-1}, \eta_t)$.

The messages $\hat{\alpha}_t(\eta_t)$ and $\hat{\beta}_t(\eta_t)$ are the approximated forms of $\alpha_t(\eta_t)$ and $\beta_t(\eta_t)$, respectively. Combining Equations (2.15), (2.16), and (3.12), the smoothing latent variable can be estimated by

$$p(\eta_t|\mathbf{Z}_{1:T}) \approx q_t(\eta_t) \propto \hat{\alpha}_t(\eta_t)\hat{\beta}_t(\eta_t) \tag{3.15}$$

$$\begin{aligned} p(\eta_{t-1}, \eta_t|\mathbf{Z}_{1:T}) &\approx \hat{p}_t(\eta_{t-1}, \eta_t), \\ &\propto \hat{\alpha}_{t-1}(\eta_{t-1})p(\eta_t|\eta_{t-1})p(\mathbf{Z}_t|\eta_t)\hat{\beta}_t(\eta_t) \\ &= \hat{\alpha}_{t-1}(\eta_{t-1})\Omega_t(\eta_{t-1}, \eta_t)\hat{\beta}_t(\eta_t). \end{aligned} \tag{3.16}$$

Furthermore, given that from the factorial form,

$$\hat{p}_t(\eta_{t-1}, \eta_t) = q_{t-1}(\eta_{t-1})q_t(\eta_t), \tag{3.17}$$

plugging Equations (3.14), (3.15), (3.16) into Equation (3.17) leads to the simplified approximation form:

$$\hat{\Omega}_t(\eta_{t-1}, \eta_t) = \hat{\beta}_{t-1}(\eta_{t-1})\hat{\alpha}_t(\eta_t). \tag{3.18}$$

The EP algorithm refines the approximate posterior $q(\eta)$ iteratively by recomputing passing messages. As indicated in Equation (3.18), in order to estimate the approximate factor $\hat{\Omega}_t^{new}(\eta_{t-1}, \eta_t)$, we need to estimate $\hat{\beta}_{t-1}^{new}(\eta_{t-1})$ and $\hat{\alpha}_t^{new}(\eta_t)$. One-slice posterior distribution can be acquired by integrating one latent variable from two-slice posterior distribution. When we compute the one-slice posterior, the corresponding message can be calculated by Equation (3.15). Hence, by combining Equations (3.16) and (3.17), these two messages can be obtained by following two steps: 1) approximating $\hat{p}_t(\eta_{t-1}, \eta_t) \propto \hat{\alpha}_{t-1}(\eta_{t-1})p(\eta_t|\eta_{t-1})p(\mathbf{Z}_t|\eta_t)\hat{\beta}_t(\eta_t)$ as a Gaussian distribution by Laplace Approximation

$$\hat{p}_t(\eta_{t-1}, \eta_t) \approx_{LA} \mathcal{N}(\eta_{t-1}, \eta_t \mid \mu, \mathbf{\Sigma}), \tag{3.19}$$

where μ and $\mathbf{\Sigma}$ match the first and second moments of $\hat{p}_t(\eta_{t-1}, \eta_t)$; 2) integrating out η_{t-1} (or η_t) to obtain $\hat{\alpha}_t^{new}(\eta_t)$ (or $\hat{\beta}_{t-1}^{new}(\eta_{t-1})$):

$$\hat{\alpha}_t^{new}(\eta_t) \propto \frac{\int \mathcal{N}(\eta_{t-1}, \eta_t \mid \mu, \mathbf{\Sigma}) d\eta_{t-1}}{\hat{\beta}_t(\eta_t)}, \tag{3.20}$$

$$\hat{\beta}_{t-1}^{new}(\eta_{t-1}) \propto \frac{\int \mathcal{N}(\eta_{t-1}, \eta_t \mid \mu, \mathbf{\Sigma}) d\eta_t}{\hat{\alpha}_{t-1}(\eta_{t-1})}. \tag{3.21}$$

The EP algorithm can be described as following steps:

1. Initialize the factors $\alpha_t(\eta_t)$ and $\beta_t(\eta_t)$ according to Equation (3.12).
2. Initialize the approximated factors $\hat{\Omega}_t(\eta_{t-1}, \eta_t)$ by $\hat{\beta}_{t-1}(\eta_{t-1})\hat{\alpha}_t(\eta_t)$.
3. Initialize the approximated posterior distribution by setting

$$q(\eta) \propto \prod_t \hat{\Omega}_t(\eta_{t-1}, \eta_t).$$

4. Until convergence:
   (a) Choose factor $\hat{\Omega}_t(\eta_{t-1}, \eta_t)$ for refinement
   (b) Remove $\hat{\Omega}_t(\eta_{t-1}, \eta_t)$ from the posterior by division

$$q^{\setminus t}(\eta) = \frac{q(\eta)}{\hat{\Omega}_t(\eta_{t-1}, \eta_t)}.$$

(c) Update $\hat{\beta}_{t-1}^{new}(\eta_{t-1})$ and $\hat{\alpha}_t^{new}(\eta_t)$ by Equations (3.20) and (3.21) based on the

$$\hat{p}_t(\eta_{t-1}, \eta_t) = \hat{\alpha}_{t-1}(\eta_{t-1}) p(\eta_t|\eta_{t-1}) p(\mathbf{Z}_t|\eta_t) \hat{\beta}_t(\eta_t).$$

(d) Update $\hat{\Omega}_T^{new}(\eta_{t-1}, \eta_t)$ by $\hat{\beta}_{t-1}^{new}(\eta_{t-1}) \hat{\alpha}_t^{new}(\eta_t)$ based on Equation (3.18)

(e) Update the posterior distribution

$$q^{new} \propto q^{\setminus t}(\eta)\ \hat{\Omega}_t^{new}(\eta_{t-1}\eta_t).$$

The above algorithm outputs the estimated messages $\hat{\alpha}_t(\eta_t)$ and $\hat{\beta}_t(\eta_t)$, $t = 1, \cdots, T$, each of which follows a Gaussian distribution, with known parameters. The posterior distributions of $p(\eta_t|\mathbf{Z}_{1:t})$, $p(\eta_t|\mathbf{Z}_{1:T})$ can be estimated as

$$\begin{aligned} \hat{p}(\eta_t|\mathbf{Z}_{1:t}) &= \frac{1}{\mathcal{Z}_{1:t}} \hat{\alpha}_t(\eta_t), \\ \hat{p}(\eta_t|\mathbf{Z}_{1:T}) &= \frac{1}{\mathcal{Z}_{1:T}} \hat{\alpha}_t(\eta_t) \hat{\beta}_t(\eta_t), \end{aligned} \tag{3.22}$$

where $\mathcal{Z}_{1:t}$ and $\mathcal{Z}_{1:T}$ are the normalization factors. The mean and variance-covariance matrix $\eta_{t|T}$ and $\mathbf{P}_{t|T}$ can be estimated readily from (3.22).

3.3 Approximating $p(\xi_t|\mathbf{Z}_{1:t})$ and $p(\xi_t|\mathbf{Z}_{1:T})$

This section focuses on the approximate estimation of the posterior $p(\xi_t|\mathbf{Z}_{1:t})$ and $p(\xi_t|\mathbf{Z}_{1:T})$. The joint posterior distribution

$$\begin{aligned} p(\xi_t, \eta_t|\mathbf{Z}_{1:t}) &= \frac{1}{p(\mathbf{Z}_{1:t})} p(\xi_t, \eta_t, \mathbf{Z}_{1:t}) \\ &= \frac{1}{p(\mathbf{Z}_t|\mathbf{Z}_{1:t-1})} p(\xi_t, \eta_t, \mathbf{Z}_t|\mathbf{Z}_{1:t-1}) \\ &\propto p(\xi_t, \mathbf{Z}_t|\eta_t, \mathbf{Z}_{1:t-1}) p(\eta_t|\mathbf{Z}_{1:t-1}) \\ &= p(\xi_t, \mathbf{Z}_t|\eta_t) p(\eta_t|\mathbf{Z}_{1:t-1}) \\ &= p(\mathbf{Z}_t|\eta_t, \xi_t) p(\xi_t|\eta_t) p(\eta_t|\mathbf{Z}_{1:t-1}). \end{aligned} \tag{3.23}$$

Given $\hat{p}(\eta_{t-1}|\mathbf{Z}_{1:t-1}) \sim \mathcal{N}(\eta_{t-1|t-1}, \mathbf{P}_{t-1|t-1})$ estimated in Chapter 3.2, it follows that

$$\begin{aligned} \hat{p}(\eta_t|\mathbf{Z}_{1:t-1}) &\sim \mathcal{N}(\eta_{t|t-1}, \mathbf{P}_{t|t-1}), \\ \text{where } \eta_{t|t-1} &= \mathbf{H}_t \eta_{t-1|t-1}, \\ \mathbf{P}_{t|t-1} &= \mathbf{H}_t \mathbf{P}_{t-1|t-1} \mathbf{H}_t^T + \mathbf{U}_t. \end{aligned} \tag{3.24}$$

The posterior $p(\xi_t, \eta_t|\mathbf{Z}_{1:t})$ can be approximated as

$$\hat{p}(\xi_t, \eta_t|\mathbf{Z}_{1:t}) = p(\mathbf{Z}_t|\eta_t, \xi_t) p(\xi_t) \hat{p}(\eta_t|\mathbf{Z}_{1:t-1}). \tag{3.25}$$

So we can estimate $p(\xi_t|\mathbf{Z}_{1:t})$ by integrating η_t to get

$$\begin{aligned} p(\xi_t|\mathbf{Z}_{1:t}) &= \int p(\xi_t,\eta_t|\mathbf{Z}_t)d\eta_t \\ &= \int p(\mathbf{Z}_t|\eta_t,\xi_t)p(\xi_t)p(\eta_t|\mathbf{Z}_{1:t-1})d\eta_t \\ &\approx \int \hat{p}(\xi_t,\eta_t|\mathbf{Z}_{1:t})d\eta_t. \end{aligned} \tag{3.26}$$

Theorem 2. *When $p(\mathbf{Z}_t|\eta_t,\xi_t)$ follows a Gaussian distribution $\mathcal{N}(\mathbf{O}_t\mathbf{X}_t\beta_t+\mathbf{O}_t\mathbf{S}_t^T\eta_t+\xi_t,\sigma_\varepsilon^2\mathbf{I})$, the posterior $p(\xi_t|\mathbf{Z}_{1:t})$ and $p(\xi_t|\mathbf{Z}_{1:T})$ estimated by Equation (3.26) is equivalent to that estimated by the regular STRE model (Equations (2.7) and (2.8)).*

Proof. If $p(\mathbf{Z}_t|\eta_t,\xi_t)$ follows a Gaussian distribution, $p(\xi_t,\eta_t|\mathbf{Z}_{1:t})$ can be shown to follow a Gaussian distribution as well. From Equation (3.23), we conclude that after computing the conditional distribution of $(\mathbf{Z}_t,\eta_t,\xi_t)^T$ given $\mathbf{Z}_{1:t-1}$, which is Gaussian, the conditional distribution of $(\eta_t,\xi_t)^T$ given $\mathbf{Z}_{1:t}$ is also Gaussian [5] and can be estimated as:

$$p\left(\begin{array}{c}\eta_t\\ \xi_t\end{array}\middle|\,\mathbf{Z}_{1:t}\right)=\mathcal{N}\left(\left[\begin{array}{c}\eta_{t|t}\\ \xi_{t|t}\end{array}\right],\left[\begin{array}{cc}\mathbf{P}_{t|t} & \mathbf{Q}_{t|t}\\ \mathbf{Q}_{t|t}^T & \mathbf{R}_{t|t}\end{array}\right]\right), \tag{3.27}$$

where $\eta_{t|t}$ and $\mathbf{P}_{t|t}$ are estimated in Chapter 3.2 and

$$\begin{aligned} \xi_{t|t} &= \sigma_{\xi,t}^2\mathbf{V}_{\xi,t}\mathbf{O}_t^T(\mathbf{O}_t\mathbf{S}_t^T\mathbf{P}_{t|t-1}\mathbf{S}_t\mathbf{O}_t^T+\mathbf{D}_t)^{-1}(\mathbf{Z}_t-\mathbf{O}_t\mathbf{X}_t\beta_t-\mathbf{O}_t\mathbf{S}_t^T\eta_{t|t-1}),\\ \mathbf{R}_{t|t} &= \sigma_{\xi,t}^2\mathbf{V}_{\xi,t}-\sigma_{\xi,t}^2\mathbf{V}_{\xi,t}\mathbf{O}_t^T(\mathbf{O}_t\mathbf{S}_t^T\mathbf{P}_{t|t-1}\mathbf{S}_t\mathbf{O}_t^T+\mathbf{D}_t)^{-1}\mathbf{O}_t\mathbf{V}_{\xi,t}\sigma_{\xi,t}^2,\\ \mathbf{Q}_{t|t} &= -\mathbf{P}_{t|t-1}\mathbf{S}_t\mathbf{O}_t^T(\mathbf{O}_t\mathbf{S}_t^T\mathbf{P}_{t|t-1}\mathbf{S}_t\mathbf{O}_t^T+\mathbf{D}_t)^{-1}\mathbf{O}_t\mathbf{V}_{\xi,t}\sigma_{\xi,t}^2\\ \text{where } \mathbf{D}_t &= \sigma_{\xi,t}^2\mathbf{O}_t\mathbf{V}_{\xi,t}\mathbf{O}_t^T+\sigma_{\varepsilon,t}^2\mathbf{V}_{\varepsilon,t}. \end{aligned} \tag{3.28}$$

Comparing Equations (3.28) and (2.7) of $\xi_{t|t}, R_{t|t}$, we conclude that posteriors $p(\xi_t|\mathbf{Z}_{1:t})$ estimated by two different approaches, STRE and St-RSTP, are identical. The proof on the equivalence of estimating $p(\xi_t|\mathbf{Z}_{1:T})$ using those two approaches can be derived similarly. □

In our St-RSTP model, $p(\mathbf{Z}_t|\eta_t,\xi_t)$ follows a Student's t-distribution instead, and hence no tractable inference is available. We again apply EP to obtain an approximate distribution $\hat{p}(\xi_t,\eta_t|\mathbf{Z}_{1:t})$ as a Gaussian distribution, and then integrate out η_t to obtain $\hat{p}(\xi_t|\mathbf{Z}_{1:t})$, as showed in Equation (3.26).

The joint distribution $\hat{p}(\xi_t,\eta_t|\mathbf{Z}_{1:t})$ comprises a product of factors in the form

$$\hat{p}(\xi_t,\eta_t|\mathbf{Z}_{1:t})=\prod_{n=1}^{N_t}\left\{p(\mathbf{Z}_{tn}|\eta_t,\xi_{tn})p(\xi_{tn})\right\}\hat{p}(\eta_t|\mathbf{Z}_{1:t-1}). \tag{3.29}$$

We approximate $\hat{p}(\xi_t, \eta_t|\mathbf{Z}_{1:t})$ as a product of factors

$$q(\xi_t, \eta_t) = \prod_{n=1}^{N_t} \left\{ q_n(\xi_{tn}, \eta_t|\hat{\mu}_{tn}, \hat{\mathbf{\Sigma}}_{tn}) p(\xi_{tn}) \right\} \hat{p}(\eta_t|\mathbf{Z}_{1:t-1}), \tag{3.30}$$

where $p(\mathbf{Z}_{tn}|\eta_t, \xi_{tn})$ is approximated by the Gaussian function

$$q_n(\xi_{tn}, \eta_t|\hat{\mu}_{tn}, \hat{\mathbf{\Sigma}}_{tn}) \sim \mathcal{N}(\hat{\mu}_{tn}, \hat{\mathbf{\Sigma}}_{tn}), \tag{3.31}$$

and $\hat{\mu}_{tn}$ and $\hat{\mathbf{\Sigma}}_{tn}$ are unknown parameters to be estimated. Notice that, given the estimated $\hat{p}(\eta_t|\mathbf{Z}_{1:t-1})$, Equation (3.25) indicates that the sets of variables $\{\xi_t, \eta_t\}$ and $\{\xi_s, \eta_s\}$ are independent when $t \neq s$. Different from the EP algorithm in Chapter 3.2, which needs to propagate the messages backward and forward to the variables at different time stamps, the EP algorithm for estimating $\hat{q}(\xi_t, \eta_t)$ can be conducted separately for different time stamps. The detailed EP algorithm for estimating $p(\xi_t, \eta_t|\mathbf{Z}_{1:t})$ can be described as follows:

1. Estimate the approximate factors $\hat{p}(\eta_{t-1}|\mathbf{Z}_{1:t-1})$ by the EP algorithm proposed in Chapter 3.2. Estimate $\hat{p}(\eta_t|\mathbf{Z}_{1:t-1})$ by Equation (3.24).

2. Initialize the factors $q_n(\xi_{tn}, \eta_t|\hat{\mu}_{tn}, \hat{\mathbf{\Sigma}}_{tn})$, $n = 1, \cdots, N_t$, by setting $\hat{\mu}_{tn} = [\mathbf{0}]$ and

$$\hat{\mathbf{\Sigma}}_{tn} = \left| \begin{array}{cc} 1 & -\mathbf{S}_{tn}^T \\ -\mathbf{S}_{tn} & \mathbf{S}_{tn}\mathbf{S}_{tn}^T \end{array} \right| \sigma_\xi^2.$$

3. Until convergence (iterate on $n = 1, \cdots, N_t$):

   (a) Remove the factor $q_n(\xi_{tn}, \eta_t|\hat{\mu}_{tn}, \hat{\mathbf{\Sigma}}_{tn})$ from $q(\xi_t, \eta_t)$ by division

$$q^{\backslash n}(\xi_t, \eta_t) \propto \frac{q(\xi_t, \eta_t)}{q_n(\xi_{tn}, \eta_t|\hat{\mu}_{tn}, \hat{\mathbf{\Sigma}}_{tn})}. \tag{3.32}$$

   (b) Estimate the new posterior $q^{new}(\xi_t, \eta_t)$ by matching the first and second moments of

$$q^{\backslash n}(\xi_t, \eta_t) p(\mathbf{Z}_{tn}|\eta_t, \xi_{tn}).$$

   (c) Update the new factor

$$q^{new}(\xi_{tn}, \eta_t|\hat{\mu}_{tn}, \hat{\mathbf{\Sigma}}_{tn}) = \frac{q^{new}(\xi_t, \eta_t)}{q^{\backslash n}(\xi_t, \eta_t)}. \tag{3.33}$$

Evaluating the above EP algorithm, the number of required iterations is greater than N_t, which is the size of locations at time stamp t. For each iteration, it needs to evaluate new posterior $q^{new}(\xi_t, \eta_t)$ by setting the first and second order moments of $q^{new}(\xi_t, \eta_t)$ equal to those of $q^{\backslash n}(\xi_t, \eta_t) p(\mathbf{Z}_{tn}|\eta_t, \xi_{tn})$. One popular strategy is to apply numerical optimization,

such as Newton's method or interior point algorithm. However, notice that the total number of variables ($\{\xi_{tn}, \eta_t\}, n = 1, \cdots, N_t$) involved in the optimization process is $N_t + r$, where r is the dimension of η_t. As we know, the performance of numerical optimization deteriorates significantly when the number of variables is large. In practice, we consistently observed that this implementation runs very slow for large datasets.

In the following, we explore the special structure of the factorized forms (3.29) and (3.30), and present a fast algorithm to implement the above steps 3(b) and 3(c). The goal of these two steps is to obtain the updated factor

$$q^{new}(\xi_{tn}, \eta_t | \hat{\mu}_{tn}, \hat{\mathbf{\Sigma}}_{tn}) \sim \mathcal{N}(\hat{\mu}_{tn}, \hat{\mathbf{\Sigma}}^2_{tn}). \tag{3.34}$$

Notice that the dependency between ξ_{tn} and $\{\xi_{ts}, s \neq n\}$ is realized only through η_t, and the joint distribution of η_t and $\{\xi_{ts}, s \neq n\}$ is Gaussian. Hence, we are able to obtain the analytical form ($\tilde{q}_n(\xi_{tn}, \eta_t)$) by marginalization over $\{\xi_{ts}, s \neq n\}$:

$$\begin{aligned}
\tilde{q}_n(\xi_{tn}, \eta_t) &= \int q^{\setminus n}(\xi_{tn}, \xi_t^{\setminus n}, \eta_t) p(\mathbf{Z}_{tn} | \eta_t, \xi_{tn}) d\xi_t^{\setminus n} \\
&= p(\mathbf{Z}_{tn} | \eta_t, \xi_{tn}) p(\xi_{tn}) \hat{p}(\eta_t | \mathbf{Z}_{1:t}) \\
&\quad \times \int d\xi_t^{\setminus n} \prod_{s=1, s\neq n}^{N_t} q_s(\xi_{ts}, \eta_t | \hat{\mu}_{ts}, \hat{\mathbf{\Sigma}}_{ts}) p(\xi_{ts}) \\
&= p(\mathbf{Z}_{tn} | \eta_t, \xi_{tn}) p(\xi_{tn}) \mathcal{N}(\eta_t | \tilde{\mu}_{\eta_t}^{\setminus n}, \tilde{\mathbf{\Sigma}}_{\eta_t}^{\setminus n}),
\end{aligned} \tag{3.35}$$

where $\xi_t^{\setminus n} = \{\xi_{ts}, s \neq n\}$

$$\begin{aligned}
\tilde{\mu}_{\eta_t}^{\setminus n} &= \tilde{\mathbf{\Sigma}}_{\eta_t}^{\setminus n} \left(\sum_{s=1, s\neq n}^{N_t} \left(\mathbf{G}_{\eta_t, s}^{-1} \cdot b_{\eta_t, s} \right) + \mathbf{P}_{t|t-1}^{-1} \eta_{t|t-1} \right), \\
(\tilde{\mathbf{\Sigma}}_{\eta_t}^{\setminus n})^{-1} &= \sum_{s=1, s\neq n}^{N_t} \mathbf{G}_{\eta_t, s}^{-1} + \mathbf{P}_{t|t-1}^{-1},
\end{aligned}$$

and $b_{\eta_t, s}$ and $\mathbf{G}_{\eta_t, s}$ refer the mean and variance-covariance matrix of $q_s(\xi_{ts}, \eta_t | \hat{\mu}_{ts}, \hat{\mathbf{\Sigma}}_{ts})$ marginalized over ξ_{ts}.

Expanding Equation (3.35), we obtain

$$\begin{aligned}
\ln \tilde{q}(\xi_{tn}, \eta_t) &= -\frac{\xi_{tn}^2}{2 v_{\xi_{tn}}} - \frac{1}{2} (\eta_t - \tilde{\mu}_{\eta_t}^{\setminus n})^T (\tilde{\mathbf{\Sigma}}_{\eta_t}^{\setminus n})^{-1} (\eta_t - \tilde{\mu}_{\eta_t}^{\setminus n}) \\
&\quad -(\frac{\nu}{2} + \frac{1}{2}) \ln(1 + \frac{\varepsilon_{tn}^2}{\nu \sigma}) + const,
\end{aligned}$$

where $\varepsilon_{tn} = \mathbf{Z}_{tn} - \mathbf{O}_{tn} S_t^T \eta_t - \mathbf{1}^T \mathbf{O}_{tn} \xi_{tn}$, and $\mathbf{O}_{tn}$ refers to the n-th row of $\mathbf{O}_t$.

The next step is to approximate $\tilde{q}(\xi_{tn}, \eta_t)$ as a Gaussian form $\tilde{f}(\xi_{tn}, \eta_t)$ by matching the first and second order moments. This can be efficiently done by using iterative reweighted least squares (IRLS) [9].

After the posterior distribution $p(\xi_t|\mathbf{Z}_{1:t})$ is approximated as a Gaussian distribution, $p(\xi_t|\mathbf{Z}_{1:T})$ can be estimated by the regular STRE model as showed in Equations (2.8), where the $\eta_{t|t}$ and $\eta_{t|T}$ have been estimated in Chapter 3.2.

The complete algorithm of proposed St-RSTP is presented as follows.

Algorithm 2: St-RSTP

---

1. Estimate the mean and variance of $p(\eta_t|\mathbf{Z}_{1:t}), p(\eta_t|\mathbf{Z}_{1:T})$ by the approach in Chapter 3.2 and Equation (3.22);
2. Estimate the mean and variance of $p(\xi_t, \eta_t|\mathbf{Z}_{1:t}), p(\xi_t, \eta_t|\mathbf{Z}_{1:T})$ by the approach in Chapter 3.3 and Equation (3.35);
3. Integral out the η_t in step 2 to acquire mean and variance of $p(\xi_t|\mathbf{Z}_{1:t}), p(\xi_t|\mathbf{Z}_{1:T})$;
4. Calculate the mean and variance of latent variable Y by Equation (3.6).

Chapter 4

Experiments

This chapter evaluates the robustness and effectiveness of our proposed St-RSTP prediction algorithm compared with the regular STRE modeling algorithm. We conducted a simulation study and comprehensive experiments on two real data sets, an Aerosol Optical Depth (AOD) data set collected by NASA and a region-wide traffic volume (TV) data set collected in the City of Bellevue, WA. The following sections describes the experiment design, and illustrates the evaluation results from our filtering processes.

4.1 Experiment Design and Evaluation

4.1.1 Experiment Design

Given the raw data, we first conducted a preprocess procedure to generate original observations $\mathbf{Z}_{1:T}$ by cleaning the data set, converting the observations into a close-to-symmetric distribution, and selecting a study region. The second step was to estimate the St-RSTP parameters based on the clean data set by applying the EM estimation method proposed by [14]. The third step was to run the STRE smoothing on the clean data set to obtain the set of smoothed values $\hat{\mathbf{Y}}_{1:T}$ as the ground truth for calculating error. The fourth step was to randomly add isolated or region (cluster of) outliers into the clean data to obtain the contaminated data set $\tilde{\mathbf{Z}}_{1:T}$ (except for TV). The fifth step was to apply the STRE filtering algorithm and the proposed St-RSTP algorithm to estimate the filtered values $\mathbf{Y}_{1:T}^{(s)}$ and $\mathbf{Y}_{1:T}^{(sr)}$, respectively. The final step was to calculate the mean absolute percentage error (MAPE) and Root Mean Square Error (RMSE), by comparing $\mathbf{Y}_{1:T}^{(s)}$ and $\mathbf{Y}_{1:T}^{(sr)}$ with $\mathbf{Y}_{1:T}$.

Recall that a filtering process is to estimate the latent random variables $\mathbf{Y}_{1:T}$, given the observations $\mathbf{Z}_{1:T}$. Their relationship has the form

$$\mathbf{Z}_t = \mathbf{Y}_t + \varepsilon_t, t = 1, \cdots, T,$$

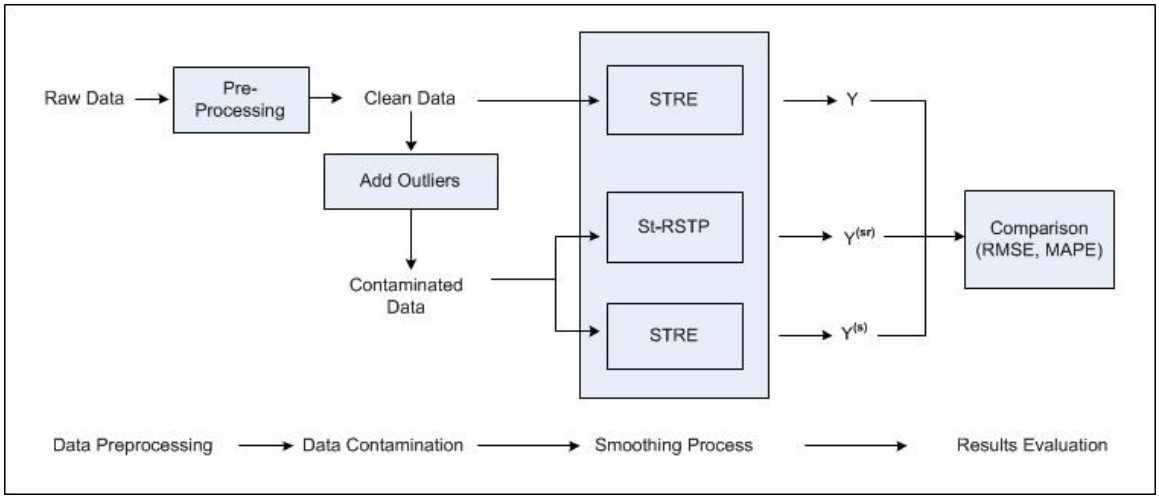


Figure 4.1: Experiment Design

where ε_t refers to the noise or the measurement error. It is observed that the filtering process is actually to estimate the unknown using surrounding denoised values.

4.1.2 Result Evaluation

We considered two widely applied measures, including MAPE and RMSE. In our simulation study, the latent random variables $\mathbf{Y}_{1:T}$ are already known. In order to evaluate the robustness, we first generated the contaminated observations $\tilde{\mathbf{Z}}_{1:T}$ by replacing a small portion of observations with some abnormal values. After that, we ran the St-RSTP and the STRE filtering algorithms on $\tilde{\mathbf{Z}}_{1:T}$ to generate the filtered values $\mathbf{Y}_{1:T}^{(s)}$ and $\mathbf{Y}_{1:T}^{(sr)}$, respectively. The MAPE measures for the regular STRE filtering and the St-RSTP filtering can be calculated as

$$\begin{aligned} MAPE^{(s)} &= \frac{1}{(\sum_{t=1}^{T} N_t)} \sum_{t=1}^{T} \sum_{n=1}^{N_t} \left\{ \frac{|Y_{tn} - Y_{tn}^{(s)}|}{Y_{tn}} \right\}, \\ MAPE^{(sr)} &= \frac{1}{(\sum_{t=1}^{T} N_t)} \sum_{t=1}^{T} \sum_{n=1}^{N_t} \left\{ \frac{|Y_{tn} - Y_{tn}^{(sr)}|}{Y_{tn}} \right\}. \end{aligned}$$

The RMSE measures for the regular STRE filtering and the St-RSTP filtering can be calculated as

$$\begin{aligned} RMSE^{(s)} &= \sqrt{\frac{1}{(\sum_{t=1}^{T} N_t)} \sum_{t=1}^{T} \sum_{n=1}^{N_t} (Y_{tn} - Y_{tn}^{(s)})^2}, \\ RMSE^{(sr)} &= \sqrt{\frac{1}{(\sum_{t=1}^{T} N_t)} \sum_{t=1}^{T} \sum_{n=1}^{N_t} (Y_{tn} - Y_{tn}^{(sr)})^2}. \end{aligned}$$

For the AOD data set, we first applied the regular STRE smoothing algorithm to the original data $\mathbf{Z}_{1:T}$ to obtain the smoothed values $\hat{\mathbf{Y}}_{1:T}$. Then we contaminated the data with a small portion of outliers and applied the regular STRE and the St-RSTP filtering algorithms to $\tilde{\mathbf{Z}}_{1:T}$ to obtain the smoothed values $\mathbf{Y}_{1:T}^{(s)}$ and $\mathbf{Y}_{1:T}^{(sr)}$. The MAPE and RMSE measures were then calculated accordingly. The robustness can be evaluated by comparing these measures. If RMSE$^{(s)}$ is larger than RMSE$^{(sr)}$ and MAPE$^{(s)}$ is larger than MAPE$^{(sr)}$, then we can conclude that our proposed algorithm is more robust than the regular STRE filtering algorithm.

In addition to the comparisons between MAPEs and RMSEs, we also visualized some representative data including the original values and the estimated values. Because the filtering process is aimed at smoothing noises and estimating unkown locations, the output values should be close to the original observations. Based on the visualization, the qualities of different filtering results (either robust or non-robust) could be easily judged visually.

Finally, we evaluated the time comparisons between STRE and proposed St-RSTP methods on simulation study dataset and AOD dataset. In this comparison, we presents the time cost of different amount of location and time stamps.

4.2 Simulation Study

This section presents a simulation study on the robustness of the proposed St-RSTP prediction algorithm, compared with that of the regular STRE filtering algorithm. In this work, we considered the same simulation model as employed in recent STRE related papers [5, 13, 14] to generate spatio-temporal simulation data.

4.2.1 Simulation Setup

The spatial domain was designed with one dimension and had the observation locations, $D = \{s : s = 1, \cdots, 256\}$. The time domain had the observation time stamps $t = 1, 2, \cdots, 50$. We assumed that the trend component $\mu(\mathbf{s}; t)$ equaled to zero and simulated the processes $\mathbf{Y}(\mathbf{s}; t)$

and $\mathbf{Z}(\mathbf{s};t)$ according to Equations (2.2) and (2.3). We assumed a stationary process with $\mathbf{S}_t = \mathbf{S}$, $\mathbf{H}_t = \mathbf{H}$, and $\mathbf{U}_t = \mathbf{U}$. The small-scale (autoregressive) process $\{\eta_t\}$ was generated by the matrix parameters $\mathbf{H}$ and $\mathbf{U}$. The spatial basis functions $\mathbf{S}$ was defined by 30 W-wavelets from the first four resolutions [21, 24, 15]. The matrix parameters $\mathbf{K}_0$, $\mathbf{H}$, and $\mathbf{U}$ were chosen to match an exponential variance-covariance $\mathbf{\Sigma}^0$, with $\mathbf{\Sigma}^0_{ij} = exp(-|i-j|/\theta).\theta =$ 25 The matrix $\mathbf{\Sigma}$ was used to calibrate the spatial dependence between the 256 locations. Specifically, $\mathbf{K}$ was obtained by minimizing the Frobenius norm distance

$$\underset{\mathbf{K}}{\text{minimize}} \|\mathbf{SKS}^T - \mathbf{\Sigma}^0\|^2.$$

The temporal dependence was calibrated by the matrix $\mathbf{H}$. $\mathbf{H}$ was obtained by minimizing the Frobenius norm

$$\underset{\mathbf{H}}{\text{minimize}} \|\mathbf{SHKS}^T - diag(\mathbf{SKS}^T)^{1/2}\mathbf{T}^0 diag(\mathbf{SKS}^T)^{1/2}\|^2,$$

where $\mathbf{T}^0$ is defined by

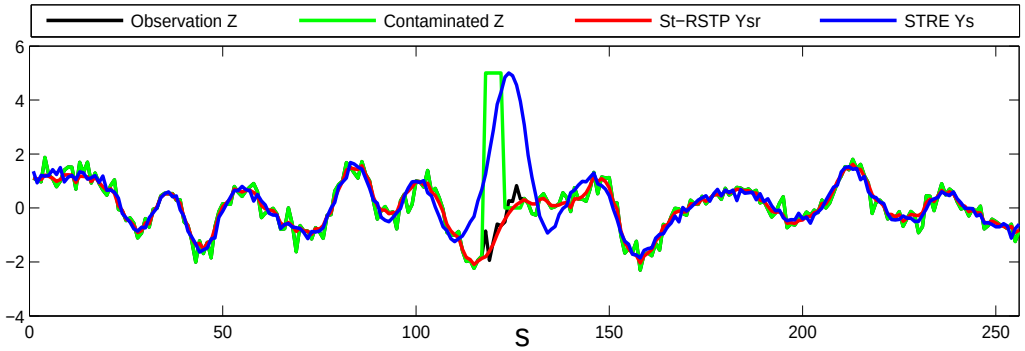
$$\mathbf{T}^0_{ij} = \rho \cdot exp\{\frac{-|i-j|}{\theta}\}, i, j = 1, ..., 256.$$

The innovation variance matrix $\mathbf{U}$ was obtained from the relationship between $\mathbf{K}, \mathbf{H}$, and $\mathbf{U} : \mathbf{U} = \mathbf{K} - \mathbf{HKH^T}$, where we always checked that $\mathbf{U}$ was positive definite. All other parameters were defined as same as the original STRE paper.

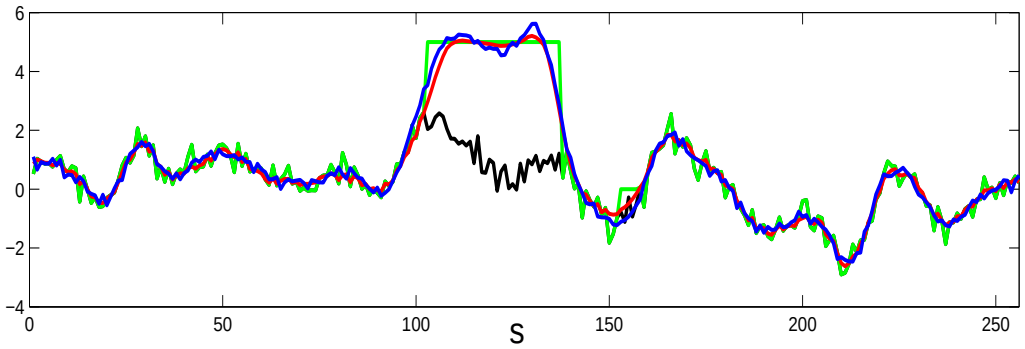
In this simulation, we tried to simulate the general real world scenarios. We considered two types of outliers namely, random outliers and regional (cluster of) outliers. For random outliers, we simulated the random outliers were brought during the data collection, which show randomly distributed in the data sets.So, we randomly picked locations and time stamps, and then shifted the observation to a larger value 5. We generated cases with 5, 15, and 35 random outliers. For regional outliers, such as sensors compromised during operating. So we fixed the center of the region and set region sizes (number of outliers) to 5, and 35. The temporal dimension of the region was fixed to a six units window. Note that, other combinations of time and spatial locations had also been tested and similar patterns were observed.

4.2.2 Simulation Results

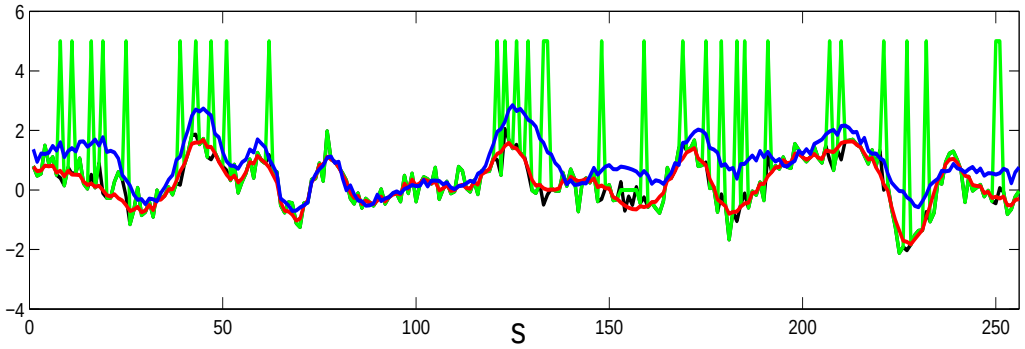
Figure 4.2 illustrates impacts of isolated outliers on the filtering algorithms at three different timestamps with various number of outliers. Each sub-figure has four curves that are related to the original observations $\mathbf{Z}_t$, the contaminated observations $\tilde{\mathbf{Z}}_t$, the filtered values $\mathbf{Y}^{(s)}$ via the regular STRE algorithm, and the filtered values $\mathbf{Y}^{(sr)}$ via our proposed St-RSTP algorithm, respectively. The X-axis refers to location index, with totally 256 distinct locations. The Y-axis denotes the $\mathbf{Z}$ or $\mathbf{Y}$ values. The symbol t refers to time stamp. As shown in the



(a) t = 4, 5 regional outliers



(b) t = 7, 35 regional outliers



(c) t = 44, 35 random outliers

Figure 4.2: STRE filtering vs. St-RSTP filtering using simulation data
(X-axis: location, Y-axis: Z or Y value)

figures, with the increasing number of outliers, the STRE curve was clearly distorted at an increasing degree. On the contrary, our proposed robust filtering algorithm demonstrated strong resilience to outlier effects. Even in the situation of high rate contaminations (35 isolated outliers, around 13% percentage in Figure 4.2(c)), our proposed algorithm could still recover the latent random variables $\mathbf{Y}_t$ very well.

Figure 4.2(a) and (b) illustrates the impacts of regional outliers on the two filtering algorithms with different outlier region sizes. When the outlier region size was small (5 adjacent outliers), our proposed robust filtering algorithm performed very well, whereas the STRE filtering algorithm was already misguided by the outliers and the filtered curve segment around the outlier region was clearly distorted. On the other hand, outside the outlier region, the filtered curve generated by St-RSTP was almost identical to the filtered curve generated from the STRE filtering algorithm. This indicates that when there are no outliers, our algorithm performs similarly as the regular STRE model, but when outliers appeared, our algorithm tends to be more resilient to the outliers.

However, we also observed that large regional outliers have significant impacts on both the STRE and St-RSTP. When we increased the region size to 35, both St-RSTP and STRE filtering algorithms were misguided and the filtered values around the outlier region were close to outlier values. This could be interpreted by the STRE model assumptions (See Chapter 2.1) that define spatio-temporal dependencies between $Z(\mathbf{s}_i; u)$ and $Z(\mathbf{s}_j; t)$, with $i \neq j$ or $u \neq t$. Particularly, the STRE model assumes a Markov Gaussian process to model spatial dependencies between $Z(\mathbf{s}_i; t)$ and $Z(\mathbf{s}_j; t)$, $i \neq j$. Observations will have a high spatial correlation if they are spatially close. For temporal dependency, the STRE model assumes a first order Markov process. That is, except for the dependence on the other locations at the current time t, $Z(\mathbf{s}; t)$ is also dependent on its previous time stamp observations $\mathbf{Z}_{t-1}$. The STRE model considers spatial Gaussian process, lag-1 temporal autocorrelation, and white noise (Gaussian distribution) to model the whole data variation. Our St-RSTP model is similar to the STRE model from this aspect except that we consider heavy tailed Student's t-distribution to model the measurement error.

Spatio-temporal outliers can be interpreted as the observations that have low correlations with their spatio-temporal neighbors and can not be regarded as the normal measurement error (white noise). When a data set has outliers, for the standard STRE model the additional variations due to outliers will be captured by distorting the spatio-temporal dependencies. The white noise component can not handle large deviations due to the non-heavy tail distribution characteristics. This explains the distorted STRE curves as shown in Figures 4.2. A specific spatio-temporal autocorrelation pattern is associated with certain degree of sharpness of the resulting filtered curves. In comparison, the St-RSTP model uses Student's t-distributions to model white noise (or the measurement error). When outliers appear, our St-RSTP model directly captures the additional large variations due to outliers as white noise. When the outlier region becomes large, however, it becomes possible to directly use the spatio-temporal autocorrelations to capture the outlier variations. Intuitively, we are able to use a smooth curve to fit the observations well. This potentially explains why the St-RSTP model could not recover the true Y values around the outlier region, when the outlier region size was large.

Table 4.1 illustrates the robustness of the filtering algorithms based on different settings of outliers. In this table, (O) refers to outliers, and (R) refers to non-outliers. It can be observed that St-RSTP algorithm always outperformed the STRE filtering algorithm in

Table 4.1: Model Robustness Comparison using Different Simulation Settings

Outlier Type	Size	$\text{MAPE}^{(sr)}$ (O)	$\text{MAPE}^{(s)}$ (O)	$\text{RMSE}^{(sr)}$ (O)	$\text{RMSE}^{(s)}$ (O)	$\text{MAPE}^{(sr)}$ (R)	$\text{MAPE}^{(s)}$ (R)	$\text{RMSE}^{(sr)}$ (R)	$\text{RMSE}^{(s)}$ (R)
Isolated Outliers	5	1.25	2.12	0.35	0.71	6.57	10.65	0.24	0.33
	15	1.34	4.89	0.33	0.84	6.62	20.06	0.24	0.35
	35	1.69	7.72	0.34	1.24	6.72	11.33	0.24	0.40
Regional Outliers	5	2.19	14.04	0.54	3.44	6.64	11.05	0.25	0.39
	35	132.14	138.94	4.58	4.75	7.22	10.82	0.25	0.33

all the scenarios we have experimented. Although we observed the similar results for 1-step forecasting, we only present the forecasting results for the real data sets due to space limitation.

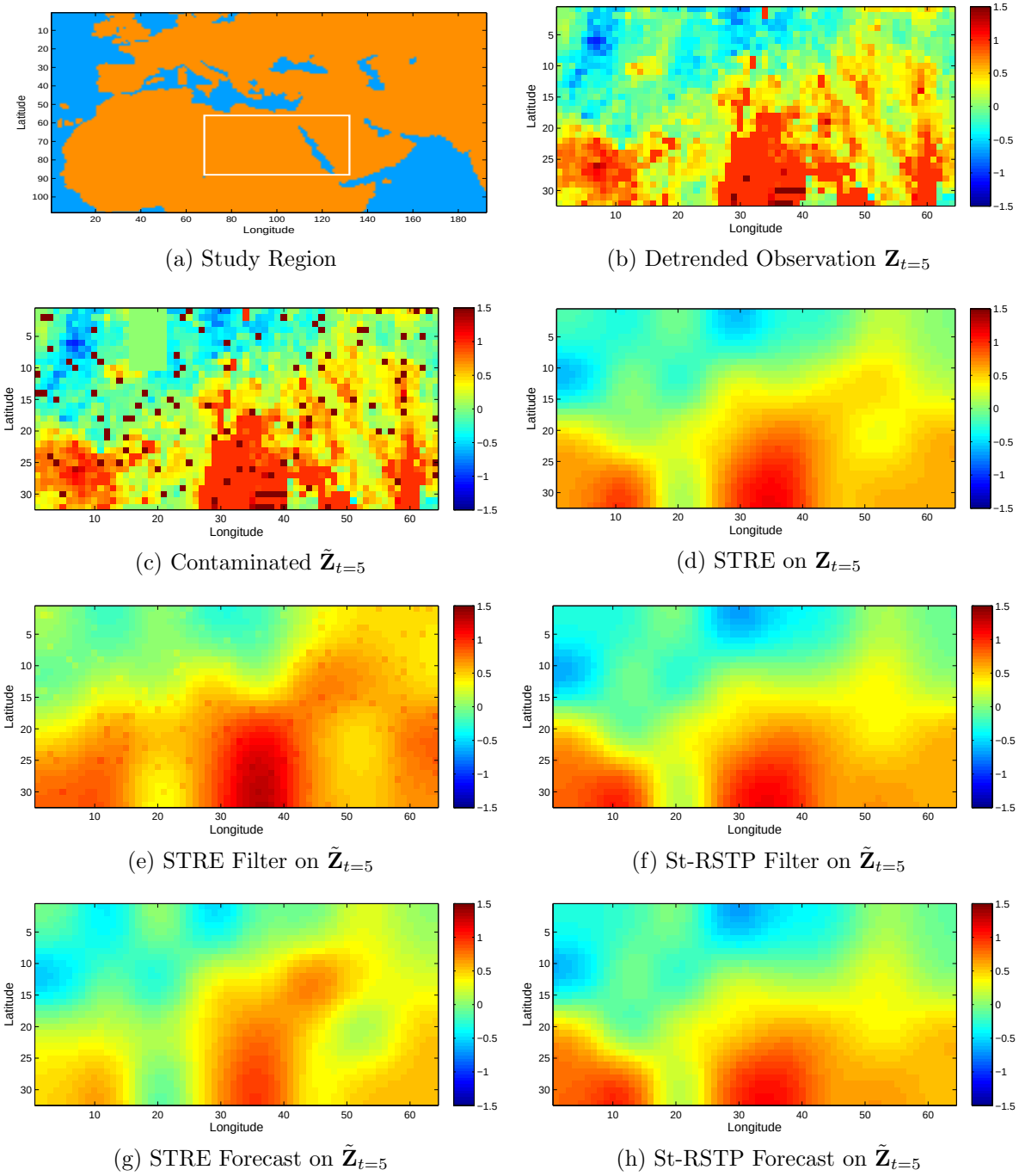
4.3 Aerosol Optical Depth Data Experiments

The AOD data set was collected by NASA's Terra satellite with MISR (Multi-angle Imaging Spectro Radiometer) on board between July 1 and August 9, 2001. Because the AOD data are heavily right-skewed, we applied log transformation $\log(AOD)$ to convert the 40-day level-3 data (with spatial resolution ($0.5^\circ \times 0.5^\circ$) and temporal resolution (1 day)) into a close-to symmetric distribution. The time unit we have chosen is eight days, and we obtain an individual datum by taking a weighted average of daily level-3 $\log(AOD)$ values in a given 8-day period, where the weight is defined by the number of level-2 observations in each level-3 pixel on each day. Time unit 1 corresponds to July 1-8, time unit 2 corresponds to July 9-16, $\cdots$, and time unit 5 corresponds to August 2-9, 2001.

We focus on the data collected in a rectangle region D between longitudes 14° and 46° and between latitudes 14° and 30°, which is part of Egypt, Saudi Arabia, as shown in Figure 4.3(a). The number of level-3 observations (pixels) in the region is $32 \times 64 = 2048$. Other geographical regions had also been studied and similar patterns were obtained.

In order to evaluate the robustness of different filtering and forecasting algorithms on the AOD data, we randomly set 5% locations in every timestamp and replaced the observations with value 5, which is outside the normal range of the observations (-0.0843 ± 0.4958).

A similar STRE model specification as used in [5] was applied in this simulation. We detrended the observations $\mathbf{Z}_t$ by the residuals $\mathbf{Z}_t - \mathbf{X}_t\beta$ to $\mathbf{Z}_t$. After this process, the observations $\mathbf{Z}$ no longer had trend components and could be called as detrended observations. The unknown parameters σ_ε^2, $\mathbf{K}_1$, and $\{\mathbf{H}_t, \mathbf{U}_t\}, t = 1, \cdots, 5$, in basis functions $\mathbf{S}$ were estimated



(a) Study Region

(b) Detrended Observation $\mathbf{Z}_{t=5}$

(c) Contaminated $\tilde{\mathbf{Z}}_{t=5}$

(d) STRE on $\mathbf{Z}_{t=5}$

(e) STRE Filter on $\tilde{\mathbf{Z}}_{t=5}$

(f) St-RSTP Filter on $\tilde{\mathbf{Z}}_{t=5}$

(g) STRE Forecast on $\tilde{\mathbf{Z}}_{t=5}$

(h) St-RSTP Forecast on $\tilde{\mathbf{Z}}_{t=5}$

Figure 4.3: STRE vs. St-RSTP on AOD data sets at time unit 5

by using the EM estimation algorithm proposed by [14].

Figures 4.3 illustrate the robustness of our St-RSTP filtering and forecasting algorithms

compared with that of the regular STRE algorithms at timestamp $t = 5$. Figure 4.3(a) shows our study region, which was within the white box on the map. Figure 4.3(b) shows the heatmap of the detrended observations $\mathbf{Z}_{t=5}$. Figure 4.3(c) displays the contaminated observations $\tilde{\mathbf{Z}}_{t=5}$, in which we injected an red-color outlier dots in the image. Figure 4.3(d) shows the STRE filtering results on the clean detrended observations $\mathbf{Z}_{t=5}$, and Figure 4.3(e) displays the STRE filtering results on the contaminated observations $\tilde{\mathbf{Z}}_{t=5}$. Figure 4.3(f) shows the St-RSTP filtering results on $\tilde{\mathbf{Z}}_{t=5}$. By comparing Figure 4.3(e) and (f) with the original filtering results shown in Figure 4.3(d), we can observe that the regular STRE filtering results were clearly distorted by the region outliers round the neighborhood area. However, our St-RSTP filtering results in Figure 4.3(f) were still very close to the original filtering results in Figure 4.3(e). Similarly, the 1-step forecasting results in Figure 4.3(g) and 4.3(h) showed that the St-RSTP produced more accurate prediction than the STRE.

Table 4.2: Model Robustness Comparison use the AOD Data

	MAPE (O)	MAPE (R)	MAPE (A)	RMSE (O)	RMSE (R)	RMSE (A)
STRE	4.3037	3.9250	3.9303	1.1420	0.3761	0.3972
St-RSTP	1.0515	2.3161	2.2983	0.4002	0.3220	0.3232

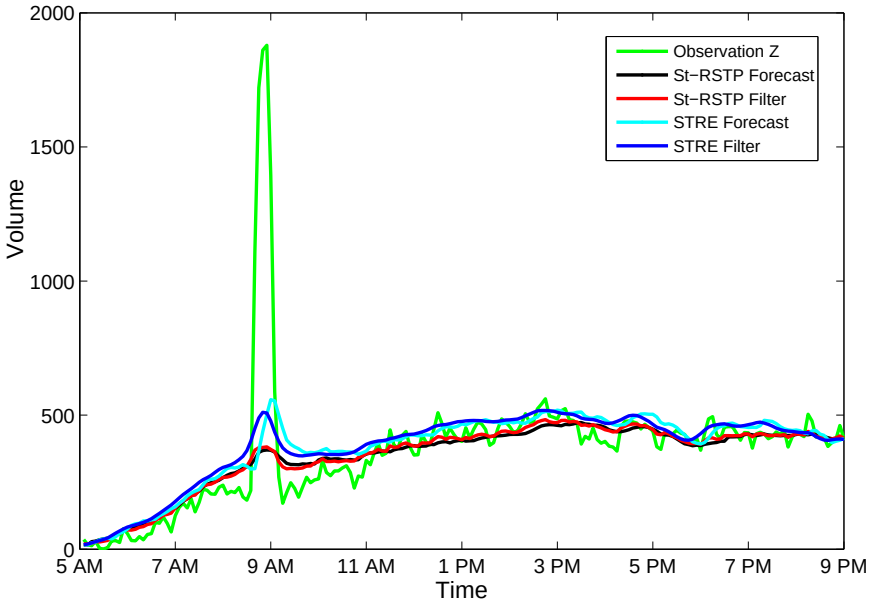
To demonstrate the results in a more comprehensive way, Table 4.2 presents the average results on all the five time units, where (O) refers to outlier region, (R) refers to non-outlier region, and (A) refers to all the region. It can be clearly observed that the St-RSTP achieved much lower MAPE and RMSE than the STRE filtering algorithm in both outlier and nonoutlier regions.

4.4 Case Study on Traffic Volume Data

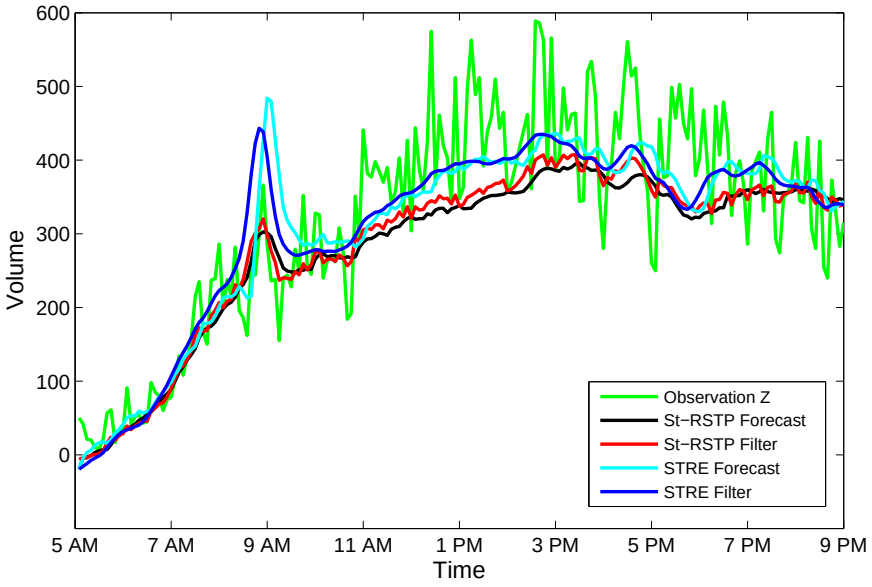
The traffic volume data were collected in the City of Bellevue, WA. The traffic volume data are collected from the advance loop detector, which is located 100 $\sim$ 130 feet (30.5 $\sim$ 39.7 m) upstream from the stop bar at each approach. As of July 2010, the City has more than 182 signalized intersections, 165 of which are controlled by traffic management center (TMC). Data from 706 loop detectors are sent to the TMC every minute. The data is currently managed by the Digital Roadway Interactive Visualization and Evaluation (DRIVE) Net system at the Smart Transportation Application and Research Laboratory (STAR Lab) [28, 17] at the University of Washington, Seattle.

In this set of experiments, 8th Ave was selected as the test route because it is a fairly busy street, with annual average weekday traffic of 37,700 (veh/day), connecting freeway I-405 and a large shopping mall (Bellevue Square). 17 detectors on NE 8th Ave are used to examine the proposed model's capability.

Weekday data (Tuesday, Wednesday and Thursday) collected from first two weeks of July, 2007 were used for training and the last two weeks of June, 2007 were used for cross validation. The verification data were collected during the first week of July in 2008. In this study, all data were aggregated into 5-minute intervals to reduce the effect of random noise. In total, the detector data collected on 17 detectors within 5376 time intervals were evaluated.



(a) t = 5th day, detector #3



(b) t = 5th day, detector #16

Figure 4.4: STRE vs. St-RSTP using the TV data on 5th day

Figure 4.4 shows the results of comparison on two detectors with different real-world outlier rates. The X-axis refers to the 192 timestamps from 5 am to 9 pm, and the Y-axis refers to the traffic volume, aggregated at 5 minute intervals. Figure 4.4(a) shows the traffic volume from detector #3 with one significant spike reached 1900 around 9 am, which was probably caused by malfunctioning. On this detector, the STRE filtering algorithm had a spike over 500 triggered by the outlier, and its 1-step forecasting had a even higher spike right after the real one. On the other hand, the St-RSTP smoothed the spike to around 300, which is closer to their spatial neighbors. The St-RSTP 1-step prediction produced the volumes very similar to its smoothed curve. Figure 4.4(b) shows the results on the detector 16 with vibrating volumes throughout the day. Because this detector was located close to detector #3 on the same route, the outlier on the detector #3 affected the STRE process on detector #16. As we can see from the figure, the STRE approach had a significant spike on the filtering curve at exactly the same time when the outlier appeared on detector #3; and a higher spike on the forecasting curve right after the outlier appeared. On the contrary, although the St-RSTP did filtering and forecasting by considering spatial and temporal neighbors as well, its process successfully resisted the impact from the spatially neighboring outlier. Besides that, one can also notice that the St-RSTP handled the vibrations on the original volume more smoothly than the STRE. More specifically, the St-RSTP forecasting gave smoother volumes than its filtering. This suggested that both St-RSTP filtering and forecasting are robust on the temporal domain. These patterns are consistent with what we observed from the simulation study and the AOD results.

4.5 Time Complexity

Table 4.3: Comparison of Time Cost using the Simulated and AOD Data

Dataset		Outliers (#)	STRE (Sec)	St-RSTP (Sec)
Simulation Data	Isolated Outliers	5	2.95	29.10
		15	3.03	29.19
		35	3.14	29.64
	Regional Outliers	5	2.72	28.28
		15	2.87	28.54
		35	2.88	28.28
AOD Data		5%	69.07	26.58

Note: The simulated data has 256 locations and 50 time units.
The AOD data has 2048 locations and 5 time units.

Table 4.3 presents execution time comparisons between our St-RSTP model and regular STRE model. The comparisons are under Windows 7 Professional 64-bit operating system, Intel core i7-Q740, 1.73GHz (CPU), 8.00 GB (RAM). We compare all the scenarios in sim-

ulation data and the whole set in AOD data. The result shows that the St-RSTP can reach ten times in execution time comparing to that of STRE algorithms under all tested simulation data scenarios. But in the AOD dataset, St-RSTP outperformed the regular STRE algorithms in all 5 time units. Our St-RSTP algorithm estimated small-scale and micro-scale variation separately. The estimation went through all the timestamps one by one, so it would cost less time and outperform STRE in a dataset with fewer timestamps.

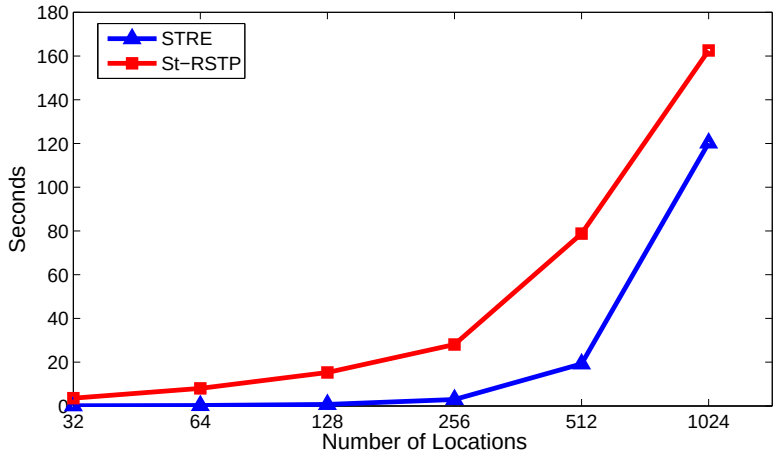


Figure 4.5: Time Cost vs. Number of Locations

On the other hand, time costs of St-RSTP and STRE on the simulation data with various location sizes are illustrated in Figure 4.5, where the X-axis shows the number of locations in log scale, and the Y-axis represents the execution time in seconds. As can be clearly observed, both St-RSTP and STRE had increased time costs when the number of locations grew up. Although the St-RSTP took longer to execute when the number of locations changed from 32 to 1024, the St-RSTP has shown better scalability than the STRE as the time differences reduced from tens of times to about 30%.

Chapter 5

Conclusion

This thesis proposes a robust and effective design of spatio-temporal prediction based on Student's t-distribution, St-RSTP. This prediction model inherits the ability of processing large scale spatio-temporal data with linear time complexity from spatio-temporal random effect model, and provides enhanced tolerance to outliers or other small departures. On the other hand, St-RSTP is applicable to provide estimation based on observations over spatio-temporal neighbors.

In the proposed design, the St-RSTP model assumes that the measurement error follows Student's t-distribution, instead of a traditional Gaussian distribution. An approximate inference algorithm in the framework of Expectation Propagation is applied to support the analytical intractable inference of Student's t model. The robustness and the efficiency of our Student-t based prediction model have been demonstrated in extensive experiments evaluations based on both simulation and real-life data sets. The proposed approach provides critical functionality for stochastic processes on spatio-temporal data.

Bibliography

[1] B. Anderson. *Adaptive Control*. Oxford: Pergamon Press, 1984.

[2] K. Arrigo, G. Dijken, and S. Bushinsky. Primary production in the southern ocean, 1997-2006. *Journal of Geophysical Research*, (113:C08004), 2008.

[3] V. Berrocal, A. Gelfand, and D. Holland. A spatio-temporal downscaler for output from numerical models. *Journal of Agrecultural, Biological, and Environmental Statistics*, 15:176–197, 2010.

[4] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[5] N. Cressie, T. Shi, and E. L.Kang. Fixed rank filtering for spatial-temporal data. *Journal of Computational and Graphical Statistics*, 19(3):724–745, 2010.

[6] N. Cressie and C. Wikle. Space-time kalman filter. *Encyclopedia of Environmetrics*, 4:2045–2049, 2002.

[7] N. Cressie and C. Wikle. *Statistics for Spatio-Temporal Data*. Wiley, 2011. ISBN 978-0471692744.

[8] S. Ghosh, P. Bhave, J. Davis, and H. Lee. Spatio-temporal analysis of total nitrate concentrations using dynamic statistical models. *Journal of the American Statistical Association*, 105:538–551, 2010.

[9] P. J. Green. Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and some Robust and Resistant Alternatives. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(2), 1984.

[10] H. Huang and N. Cressie. Spatio-temporal prediction of snow water equivalent using the kalman filter. *Computational Statistics and Data Analysis*, 22:159–175, 1996.

[11] G. Johannesson, N. Cressie, and H. Huang. Dyanmic multi-resolution spatial models. *Environmental and Ecological Statistics*, 14:5–25, 2007.

[12] Kalman, Rudolph, and Emil. A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

[13] E. Kang, N. Cressie, and T. Shi. Using temporal variability to improve spatial mapping with application to satellite data. *Canadian Journal of Statistics*, 38:271–289, 2010.

[14] M. Katzfuss and N. Cressie. Spatio-temporal smoothing and em estimation for massive remote-sensing data sets. *Journal of Time Series Analysis*, 32(4):430–446, 2010.

[15] M. Kwong and P. Tang. W-matrices, nonorthogonal multiresolution analysis, and finite signals of arbitrary length. *Technical Report MCS-P449-0794*, 1994.

[16] H. Lopes, E. Salazar, and D. gamerman. Spatial dynamic factor analysis. *Bayesian Analysis*, 3:759–792, 2009.

[17] X. Ma, Y. Wu, and Y. Wang. Drive net: An e-science of transportation platform for data sharing, visualization, modeling, and analysis. *Transportation Research Record: Journal of the Transportation Research Board*, 2215:37–49, 2011.

[18] K. Mardia, C. Goodall, E. Redfern, and F. Alonso. The kriged kalman filter. *Environmental and Ecological Statistics*, 14:5–25, 1998.

[19] T. P. Minka. From hidden markov models to linear dynamical systems. Technical report, Tech. Rep. 531, Vision and Modeling Group of Media Lab, MIT, 1999.

[20] T. P. Minka. Expectation propagation for approximate bayesian inference. In *UAI*, pages 362–369, 2001.

[21] D. Nychka, C. Wikle, and J. Royle. Multiresolution models for nonstationary spatial covariance functions. *Statistical Modeling*, 2:315–331, 2002.

[22] C. Park, W. Bridewell, and P. Langley. Integrated systems for inducing spatio-temporal process models. In M. Fox and D. Poole, editors, *AAAI*. AAAI Press, 2010.

[23] A. V. Pasi Jylanki, Jarno Vanhatalo. Robust gaussian process regression with a student-t likelihood. *Journal of Machine Learning Research*, 12:3227 – 3257, 2011.

[24] T. Shi and N. Cressie. Global statistical analysis of misr aerosol data: A massive data product from nasa's terra satellite. *Environmetrics*, 18:665–680, 2007.

[25] R. H. Shumway and D. S. Stoffer. *Time Series Analysis and Its Applications With R Examples*. Springer, 2006. ISBN 978-0-387-29317-2.

[26] C. Wikle and L. Berliner. A bayesian tutorial for data assimilation. *Physica D: Nonlinear Phenomena*, 230(1-2):1–16, 2007.

[27] C. Wikle and N. Cressie. A dimension-reduced approach to space-time kalman filtering. *Biometrika*, 86:815–829, 1999.

[28] Y.-J. Wu, S. An, X. Ma, and Y. Wang. Development of a web-based arterial network analysis system for real-time decision support. *Transportation Research Record: Journal of the Transportation Research Board*, 2215:24–36, 2011.