PREDICTING EMOTIONS FROM TEMPORAL PHYSICAL AND
BEHAVIORAL INFORMATION

A Thesis in
Industrial Engineering
by
Suppawong Tuarob

© 2015 Suppawong Tuarob

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

August 2015
The thesis of Suppawong Tuarob was reviewed and approved* by the following:

Conrad S. Tucker  
Assistant Professor of Industrial and Manufacturing Engineering  
Thesis Advisor

Soundar Kumara  
Allen E. Pearce/Allen M. Pearce Professor of Industrial and Manufacturing Engineering  
Professor of Computer Science and Engineering  
Thesis Reader

David A. Nembhard  
Associate Professor of Industrial and Manufacturing Engineering  
Graduate Program Coordinator

*Signatures are on file in the Graduate School.
Abstract

It is believed that mental anomalies such as stress and anxiety not only cause suffering for the individuals, but also lead to tragedies in some extreme cases. The ability to predict the mental states of an individual could prove critical to healthcare practitioners. Currently, the practical ways to predict an individual’s mental state is through mental examinations that involve psychological experts performing the evaluations. However, such methods may be time and resource consuming, mitigating its broad applicability to a wide population. Furthermore, some individuals may also be unaware of their mental anomalies or may not feel uncomfortable to express themselves during the evaluations. Hence, their mental anomalies could remain undetected for a prolonged period of time. The hypothesis of this work is to prove that current and future mental states of an individual could be estimated using only the information directly observable from his/her physical activities and behaviors. The problem of emotion prediction is transformed into the time series forecasting problem where an individual is represented as a multivariate time series stream of monitored physical and behavioral attributes. A mathematical model is then generated to capture the dependencies among these attributes, which is used for prediction of emotion states. In particular, we first illustrate the drawbacks of traditional multivariate time series forecasting methodology such as vector autoregression. Then, we show that such issues could be mitigated by using machine learning regression techniques which are modified for capturing temporal dependencies in multivariate time series data. A case study using the data from 150 human participants from the Pennsylvania State University community illustrates that the proposed machine learning based forecasting methods are more suitable for high-dimensional psychological data than the traditional vector autoregressive model in terms of both magnitude of error and directional accuracy. These results not only present a successful usage of machine learning techniques in psychological studies, but also serve as a building block for multiple medical applications that could rely on an automated system to gauge individuals’
states of emotion.
# Table of Contents

List of Figures ........................................................... vii
List of Tables .............................................................. viii
Acknowledgments ............................................................ ix

## Chapter 1

**Introduction** ......................................................... 1
  1.1 Problem Statement ................................................. 4
  1.2 Main Contributions ................................................. 5

## Chapter 2

**Background and Related Works** .................................. 6
  2.1 Predicting and Monitoring Behavioral and Emotional Attributes . 6
  2.2 Techniques for Multivariate Time Series Prediction and Their Applications in Healthcare Domains ........................................... 7
    2.2.1 Vector Auto Regression Based Techniques ....................... 7
    2.2.2 Machine Learning Based Techniques .............................. 8

## Chapter 3

**Methodology** ......................................................... 10
  3.1 Data Acquisition .................................................. 11
  3.2 Data Preprocessing ................................................ 13
    3.2.1 Handling Different Data Types .................................. 15
    3.2.2 Handling Different Data Frequencies ........................... 15
    3.2.3 Handling Missing Values ........................................ 16
    3.2.4 Handling Discontinuous Data ................................... 16
  3.3 Feature Space Modification ........................................ 16
Chapter 4
Case Study 20
4.1 Ground Truth Validation Data ................................. 20
4.2 Selected Target Attributes ................................. 21
4.3 Implementation ........................................ 22
4.4 Forecasting Evaluation Protocol ......................... 23
4.5 Forecasting Evaluation Metrics ......................... 23
4.6 Objective 1: Model Selection ......................... 24
  4.6.1 Stage 1: Partial Comparison ....................... 24
  4.6.2 Stage 2: Comprehensive Comparison ........... 24
4.7 Objective 2: Predicting Latent Attributes .......... 27
  4.7.1 Impact of Different Sources of Information .... 28
  4.7.2 Impact of Different Lags ......................... 31

Chapter 5
Conclusion and Future Direction 33

Bibliography 33
List of Figures

3.1 The high-level diagram of the methodology. .................................. 11

4.1 Distribution of the Positive Effect ($\text{PosEffect}_d$) attributes of the 150 participants. The Y axis marks the level (0 to 100), and the X axis represents ages in years of the participants. Each line is corresponding to a participant. Each dot is represents a daily positive effect level. .................................................. 20

4.2 Illustration of the three bursts during which the data was collected. 21

4.3 (Left) Comparison of example forecasting results by RF with different horizons (i.e. $h = 0, 1, 2, 3$) against the actual values of the attribute $\text{SATLIFE}_d$ of a participant. (Right) Comparison of absolute error of each horizon (i.e. $h = 0, 1, 2, 3$). .................. 27

4.4 Comparison of average mean absolute error (MAE) produced by Random Forest forecasters trained with different information sources (i.e. OU = both observable and latent information, O = only observable information, U = only latent information) at each horizon (days ahead of prediction). Each prediction is an average of prediction using different lags (i.e. lags = 1, 3, 5, and 7). Each attribute value has a range of $[0, 100]$. ......................... 28

4.5 Comparison of directional accuracy (DAC), range between $[0, 1]$, produced by Random Forest forecaster trained with different information sources (i.e. OU = both observable and latent information, O = only observable information, U = only latent information) at each horizon (days ahead of prediction). Each prediction is an average of prediction using different lags (i.e. lags = 1, 3, 5, and 7). ......................... 30
List of Tables

3.1 Data associated with different attributes are collected at different frequencies. ............................................. 12

3.2 Sets of questionnaires, each aiming to quantify particular observable and latent characteristics of the participant, along with their references (if any). .......................................................... 14

3.3 List of forecasting models considered, along with their references. .......................................................... 18

4.1 List of select target attributes along with their references. Refer to Table 3.2 for the comprehensive description of each attribute. .......................................................... 22

4.2 Comparison of the forecasting results of the three sample target attributes on the select forecasting models with lag $l = 3$ and forecasting horizon $h = 1$. The models are trained with partial training data (i.e. using dimension-reduced data with no missing values). .......................................................... 25

4.3 Average forecasting results of all the 150 participants using full data ($L = 7$ and $H = 1$) of each some forecasting models on the select six attributes. .......................................................... 26

4.4 Comparison of the forecasting performance between Random Forest forecasters trained with data of different lags (i.e. $L = 1, 3, 5, 7$) on the select latent attributes. The predictions of each horizon (i.e. 0 to 7) are averaged. .......................................................... 31
Acknowledgments

This thesis is a result of cumulative successful collaboration among many individuals. I would not have been able to finish this thesis without people around me to give me guidance and support.

First of all, I would like to thank my advisor, Dr. Conrad Tucker, for allowing me to join his research team and work with many individuals from diverse background. He has given me chances to expand my horizon by applying my research skills on different areas of research in the industrial engineering and healthcare domains. Working with him has exposed me to a wide range of real-world research opportunities where my skill sets would be useful. His guidance and support have been a great addition to my knowledge. In addition, I would like to thank collaborators, Dr. Nilam Ram (HDFS, Penn State), Dr. Lee Giles (IST, Penn State, also my PhD advisor), and Dr. Soundar Kumara (IE, Penn State) for their guidance on this thesis. As a young scientist, mastering research profession was one of the biggest challenges in my life. Their constant effort to provide me with guidance and enthusiasm enabled me to achieve what I never had deemed possible.

I would like to thank my internship supervisors, Dr. Kevin Glass (PNNL), Dr. Line C Pouchard (Purdue), Dr. Natasha Noy (Stanford), Dr. Jeffery S Horsburgh (Utah State University), Giri Palanisamy (ORNL), Sergei Alonichau (Microsoft), and Dr. Ray Strong (IBM), for allowing me to foster and apply my research skills on real-world problems.

I consider myself fortunate for having a chance to become acquaintance with many people in State College. Their support and friendship have been valuable. Specifically, I would like to thank my lab mates: Abhinav Singh, Ajay Bharathi, Andrew Depenbusch, Bryan Dickens, Chinmay Sane, David Munoz, Derauk Gibble, Gabe Harms, Gautam Manohar, Matthew Dering, Nitish Vasudevan, Owen Shartle, Paronkasom Indradat, Peifeng Yin, Steven Sellers, Suja John, Sungchoon Lim, Sung Woo Kang, Steven Sellers, Todd Bodnar, Yixiang Han, and Zackary Ridall. In addition to exchanging technical opinions with them, their presences have made the lab fun and productive places to work.

Finally, I would like to thank my family, Sa-ngad Tuarob (Dad), Chantana
Tuarob (Mom), and Sopawan Tuarob (Sister), for their continuing support on whatever decisions I have made. They have been an ultimate source of motivation that prevented me from giving up.

It has been a great fun chunk of time at Penn State. I have been and will be cherishing every moment that I had there.
Chapter 1

Introduction

It is believed that mental anomalies may be the root causes of many tragedies and incidents [24, 69, 81]. For example, a recent plane crash was the collateral damage of a suicide incident committed by a co-pilot who was believed to have mental problems\(^1\). Lim et al. reported that 94.5\% of suicidal incidents in Korea are caused by stress or mental disorders [42]. Such suicide and suicide attempts not only cause sentimental damages to those living, but also impose financial burdens to the society. Approximately, $34.6 billion a year is spent on combined medical and work loss costs in the United States\(^2\). In addition to suicidal activities, mental anomalies can cause violence that often results in fatalities, including homicides. Neustatter described 7 different motives for murder, all of which involve mental distortion [50]. In more common scenarios, life changing events or critical incidents often leave mental strain on one’s mind that results in unhappiness or, in some cases, chronic mental diseases.

Such tragedies could be prevented if mental anomalies could be detected during the early stages of the disease, where special care and treatment could be provided. For example, intervention and careful observation could be provided by medical specialists to individuals who have high risks of causing tragic actions. Since determining what is going on in someone’s mind from their appearance or behavior is still advanced psychological science that has not been implemented in an automated fashion, most mental diagnosis solutions involve active participation of patients and medical experts [59]. Though current solutions that involve conducting screening tests exist, such solutions would not be feasible for large population due to financial and time constraints. Furthermore, a study has shown that diagnosis based methods may discourage sick individuals from participating [4]. As a result, these mental sufferers would remain undetected, or under-treated.

Oftentimes, the states of an individual’s emotion have direct impact on his/her

\(^1\)http://www.cnn.com/2015/04/02/europe/germanwings-plane-crash-main/
\(^2\)http://www.cdc.gov/violenceprevention/suicide/consequences.html
behavioral traits, and vice versa. For example, a person may experience an intense stress after losing a job, which may later cause him/her to consume extraordinary levels of alcohol. Similarly, positive interactions with friends may decrease the level of one’s stress. If correlations between mental states and physical behavior could be captured, then it would be possible to create a mathematical model that infers one’s mental states based on his/her physical, observable activities, minimizing needs for active participation in diagnosis sessions. An immediate application of such an innovation would be a patient mentality sensor that quantifies and predicts the risk of mental anomalies of a patient. It is our conjecture that an individual’s mental states can be inferred from his/her physical behaviors that can be easily monitored. A trivial, illustrative example would be the ability to predict whether someone is happy with life from his/her alcohol intake, hours of sleep, and level of socialization. The ability to model the interaction between behavioral and emotional attributes could also shed light onto multiple psychological and healthcare applications. For example, a recommendation model could be built from the history of a patient who suffers from chronic stress to suggest him/her proper actions to avoid encountering situations that may trigger emotional instability. As another example, a forecasting model could be generated to monitor one’s emotional state using his/her observable behavioral attributes as signals. Another dimension of the application would be recommendation systems that suggest products (songs, drinks, movies, etc.) based on users’ current emotions.

Existing literature has studied emotional and behavioral development at the aggregate level, where a model is developed to explain the phenomena for an entire population. To the best of our knowledge, we are the first to pursue this investigation at the individual level, where a mathematical model is built to predict and forecast emotional states for each individual. In this research, a set of mathematical models are proposed for predicting individual mental states, using the information observable from daily activities. First, we frame the problem as a multivariate time series forecasting problem, where each individual is represented with multivariate time series stream characterizing his/her daily mental and physical statuses and activities. Each multivariate time series is a set of attributes, each of which carries a temporal stream of (typically daily) values. An attribute is an individual measurable property of a phenomenon being monitored or collected. Attributes can be divided into two categories: observable and latent. An observable attribute quantifies the level of an observable physical activity or behavior. Example of observable attributes include number of hours of sleep, number of drinks, and number of friends in the recent conversation. On the contrary, a latent attribute quantifies the level of a specific dimension of emotion, such as stress, concern, anxiety, etc. While observable attributes are objective and can be easily observed by a third person (or an external sensor), latent attributes can be difficult to observed from outside without mental evaluations from psychological experts or well established self-evaluation methods (e.g., the ones used to collect ground truth validation data...
in this work). Hence, the ability to infer and predict these latent attributes from the observable information could prove to be critical to multiple psychological-related applications, especially those involving the detection of mental anomalies. Multiple time series forecasting techniques are explored, including the traditional vector autoregressive model and machine learning based models. A case study of 150 participants, whose observable and latent attributes are collected in a 60 days of data collection campaign, is used to validate and compare the efficacy of the forecasting models.

The multivariate time series forecasting problem involves learning of historical multivariate information in order to predict the future value of an attribute of interest. Though traditional statistical based techniques for multivariate time series forecasting already exist (such as vector autoregression (VAR), its periodic-aware variant Vector Autoregressive Integrated Moving Average (VARIMA), and State Space models), and are widely used in chronic psychological studies [48,56], these models are not always applicable due to the following reasons:

1. They are not well designed to handle high-dimensional time series data [17]. The multivaraite time series data used in this thesis is high-dimensional (i.e., having a large number of attributes), consisting of at least $132 \times l$ dimensions, where $l$ is the lag. Such data could introduce too many variables that not only over-consume computational resources, but also induce false relationships between attributes that may impede the forecasting performance. Though preprocessing techniques exist that reduce the dimension (e.g., PCA) or selecting a subset of attributes, to project the high-dimensional data onto a lower dimensional one, such preprocessing techniques could eliminates useful information, allowing the time series models to capture the relationships from only partial data.

2. They have various assumptions on the characteristics of the data (e.g., stationary, linear relationship, white noise only, independency among attributes, etc. [68]). While our dataset is not always well formed due to having missing data and being sparse, and the attributes are not guaranteed to be independent, these traditional multivariate time series techniques may fail to model our data.

In the past decade, multiple machine learning algorithms have been developed and optimized. Prevalent applications of machine learning include classification, regression, and clustering. Though different machine learning algorithms have different advantages and disadvantages, these algorithms are known for the ability to deal with high-dimensionality, non-linear relationships, and flexibility in datasets (e.g., missing data and different data types including string and nominal). In this research study, we propose to use machine learning regression algorithms for the multivariate time series forecasting task. A comparison study of applying machine
learning algorithms on psychological multivariate time series data which has high
dimensionality and non-linear relationships shows that the relationship between
observable and latent attributes of an individual can be accurately modeled. Fur-
thermore, we find it is possible to infer or predict the values of latent attributes
using only the observable information.

1.1 Problem Statement

Our goal is to generate a person-specific individualized mathematical model capa-
ble of predicting a certain latent attribute of an individual, using information that
can be easily and objectively observed. Such a problem is framed into the multi-
variate forecasting framework, in which a set of generalized algorithms is developed
to handle. Let $\mathbb{P} = \{p_1, p_2, ..., p_K\}$ be the entire population, where $p_k$ represents an
individual. Mathematically, an individual $p_k \in \mathbb{P}$ is represented as a sequence of $n$
attribute vectors (i.e. $n$ data points), each of which has $m$ attributes representing
the values of the attributes at a specific time period. The individual $p_k$ can be
represented with a matrix notation as:

$$
p_k = \begin{pmatrix}
a_{11} & a_{12} & a_{13} & \cdots & a_{1m} \\
a_{21} & a_{22} & a_{23} & \cdots & a_{2m} \\
a_{31} & a_{32} & a_{33} & \cdots & a_{3m} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nm}
\end{pmatrix}
$$

(1.1)

In our setting, $m$ is the number of attributes of a participant from whom the
data was collected for $n$ days. Hence, the attribute $a_{i(j)}$ is the value of the attribute
$a_i$ collected on the $j^{th}$ day.

An attribute could either be observable or latent. An observable attribute per-
tains to an objective quantity that can be observed by other human or machine
observers, such as number of today’s drinks, number of hours of exercise, number
of friends in the recent conversation, etc. A latent attribute is a subjective quan-
tity measured by either the individual him/herself or evaluated by psychological
experts. Such attributes mostly represent feelings and mental conditions, such as
how satisfied with life, how successful you feel, whether she enjoys the weather
today, etc.

An attribute can be measured numerically or categorically. A numeric attribute
is usually presented with a percentage or a natural number, while the value of
a categorical attribute has to belong to one of the predefined classes (such as
education degree, day of week, ethnicity, etc).

Given an individual $p_k$, our task is to build an individualized forecasting model
$f(p_k, a_t, n, h)$ that learns from the individual’s data (i.e. $p_k$) of the past $n$ days,
and predicts the value of the target attribute $a_{i(n+h)}$ of the next $h$ days in advance
(assuming that the current day is the $n^{th}$ day). The value $h$ is also called the horizon or forecast horizon throughout the thesis. Note that, when the attribute $a_t$ is designated as a target attribute, it is also treated as an external variable (that is, the prediction model does not learn from its historical values). The prediction at $h = 0$ means that it is made for today (current) value. Mathematically, given an individual $p_k$, associated $n$ days of historical data, and a target attribute $a_t$, we would like to find a function $f(p_k, a_t, n, h)$ that estimates the value of $a_t$ in the next $h$ days.

1.2 Main Contributions

This thesis has the following main contributions:

- We present a modification to traditional machine learning based regression algorithms so that they can be used for forecasting seasonal multivariate time series data. Specifically, the time-delay embedding algorithm is applied to the feature space as a preprocessing step to allow traditional machine learning algorithms to process the data in temporally dependent manner.

- We find the best multivariate time series model that would best captured the temporal dependencies in the data by comparing the performance between multiple machine learning algorithms from different families and the traditional VARX model (baseline).

- We demonstrate that latent attributes can be most effectively predicted using only information from the observable attributes.

The remainder of this thesis is organized as follows: Chapter 2 discusses some background of the related literature. Chapter 3 describes the proposed machine learning based time series forecasting techniques used in this thesis, along with the methods used to investigate the possibility to forecast the latent attributes. Chapter 4 discusses the case study, the ground truth validation dataset, results, and related discussion. Chapter 5 concludes the thesis.
Background and Related Works

Though extensive computational psychology literature has studied the development and relationship between multiple mental and physical attributes, most works have explained the phenomenon at the aggregate level as opposed to an individual by individual basis. Furthermore, these works also investigated the relationship (especially correlation) of a few focus attributes, while our work is more comprehensive, in the sense that we aim to capture the relationship among a large number of attributes. The ability to generate a person-specific model would enable fine grain prediction of individual emotion states, which could potentially give rise to multiple personalized emotion based applications such as monitoring and recommendation systems. In this thesis, we aim to generate a person-specific model for each individual using his/her historical observable information. In this chapter, relevant literature is discussed.

2.1 Predicting and Monitoring Behavioral and Emotional Attributes

The increase in health concern behooves the ability to monitor and predict certain health attributes so that appropriate actions or treatment can be provided in a proactive manner. Though diagnosis based methods exist that involve health practitioners evaluating the patients, efforts have been made in the literature to replace such human-based methods with automated ones to reduce costs and increase accuracy and consistency. In this work, our goal is to predict emotional attributes from various observable signals. Hence, we discuss some previous relevant research work with similar applications.

Choudhury et al. proposed a method to detect depression in Twitter [18]. A number of features are extracted from a Twitter message including engagement, ego-network, emotion, linguistic style, and user engagement. A Support Vector
Machine classifier is trained with these features to detect the level of depression in each Twitter message. However, their methodology not only is limited to social media users, but also does not take temporal dimension into account.

Litman and Forbes-Riley showed that acoustic-prosodic and lexical features can be used to automatically predict students’ emotions in computer-human tutoring dialogues [43]. They examined emotion prediction using a classification scheme developed for prior human-human tutoring studies (negative/positive/neutral), as well as using two simpler schemes proposed by other dialogue researchers (negative/non-negative, emotional/non-emotional). Their methods were developed to handle transcribed data which is different from ours. Furthermore, their method only indicates three polarities of emotions (i.e. negative, positive, and neutral), while our methods aim to quantify the level of a dimension of emotion (i.e. level of stress, anger, happiness, etc.).

Korhonen et al. presented TERVA, a system for long-term monitoring of wellness designed for home usage [34, 35]. The system runs on a laptop and is able to monitor physiological attributes such as beat-to-beat heart rate, motor activity, blood pressure, weight, body temperature, respiration, ballistocardiography, movements, and sleep stages. In addition, self-assessments of daily well-being and activities are stored by keeping a behavioral diary. The accuracy of the system was reported to be 70-91%. This work had success in monitoring some observable attributes without the supervision of human experts. Though their methodology does not involve prediction of emotions, in our future work, we could extend their system to build a monitoring system that collects the daily routines and observable behavior. This collected observable information can then be used to build a prediction model for the target latent attributes.

2.2 Techniques for Multivariate Time Series Prediction and Their Applications in Healthcare Domains

Time series forecasting techniques have been well studied and applied to multiple applications. In this thesis, the problem is framed as a multivariate time series forecasting problem; hence, in this section, related works in healthcare and biomedical informatics related to time series forecasting techniques are discussed.

2.2.1 Vector Auto Regression Based Techniques

Vector Autoregression (VAR) [74] models have been successfully used to capture linear interdependencies among multiple univariate time series, and have been shown effective in forecasting tasks in financial [33, 73], meteorology [29, 70], biomedical [31, 85] domains, etc.
A VAR model describes the evolution of a set of \( m \) attributes over the same sample period \((t = 1, ..., T)\) as a linear function of only their past values. The attributes are collected in a \( m \times 1 \) vector \( y(t) \), whose \( i^{th} \) element, \( y_i(t) \), represents the time \( t \) observation of the \( i^{th} \) attribute. For example, if the \( i^{th} \) attribute is *Number of Drinks*, then \( y_i(t) \) is the number of drinks that the individual had on day \( t \).

A \( l^{th} \) order VAR, denoted \( VAR(l) \), can be written as:

\[
y(t) = c + A_1 y(t-1) + A_2 y(t-2) + ... + A_l y(t-l) + e(t)
\]  

(2.1)

Informally, \( VAR(l) \) predicts the value of \( y(t) \) by modelling linear relationship among the attributes observed in the past \( l \) days. Furthermore, \( l \) is often referred to as the *lag* in this work.

Latif et al. employed multivariate auto-regression to model a two-channel set of electromyography (EMG) signals from the biseps and triseps muscles [40]. The coefficients of the model are used to define the direct transfer function (DTF), which later is used as a frequency domain features to train a Support Vector Machine classifier to classify an EMG into either *extension* or *flexion* classes.

All the previous works discussed in this section assume that the relationships among the attributes in the data are linear, while such an assumption is relaxed in this work. Furthermore, the data involved in such works are quite small in dimensions (less than 10), while the time series models in this work are developed to handle high-dimensional data.

2.2.2 Machine Learning Based Techniques

Computational psychology has recently become more advanced and complex, requiring observational and experimental data from multiple participants, times, and thematic scales to verify their hypotheses. This data not only grows in magnitude, but also in its dimension. While vector autoregressive models (and its variants VARX, VARIMA, VARIMAX models) have been widely used for modeling multivariate time series data, such models face the following drawbacks that prevent them from being generalized to high-dimensional, more complex data.

- **They cannot model non-linear relationships among attributes.** It has been shown in the late 1970s and early 1980s that linear models are not well adapted to many real-world applications [23]. While study has revealed that human brains can no longer be modeled with a linear model [20], these vector autoregressive models may not be suitable for investigation of psychological phenomena. While multiple computational psychology works have shown successful usage of VAR based models to capture linear relationships among attributes [7–9], these works could have missed to include necessary attributes which may exhibit non-linear relationships.
• **They have certain requirements of the data that must be met.** Darlington mentioned that certain requirements (such as completeness, stationary, and independency) in the dataset must be met so that VAR based models can be built [17]. While psychological experimental data can be noisy, unpredictable, and not always independent, these requirements are hardly satisfied. Despite the inappropriate uses of VAR based models in computational psychology works, multiple studies in various fields also discourage the use of these VAR based models in multivariate time series forecasting tasks [64,67,87].

• **They are not suitable for high-dimensional time series data.** A VAR model of $n$ attributes with the lag of $l$ needs to keep track of $(2 + l)n + ln^2$ variables. This number of variables can be handled in the case of small problems which involve only a few attributes (i.e. $n$ is small, typically fewer than 10). However, VAR models can become significantly inefficient when dealing with high-dimensional data such as ours. In the future, the system proposed in this thesis would have to handle massive data points from multiple participants. In the era of big data, where massive and heterogeneous data needs to be processed efficiently and effectively, VAR based models may not scale well in these data intensive applications.

Machine learning techniques have been widely investigated and developed in the past decade. A wide range of applications that emerge from such techniques make machine learning algorithms suitable and applicable for many problems that aim to discover knowledge from data such as clustering, classification, and regression. Recently, Bontempi proposed extensions to machine learning algorithms to add the capability to model time series dependencies [10]. However, their methods only handle univariate time series data. Hegger et al. proposed the time-delay embedding technique which modifies the traditional feature space of machine learning algorithms so that history of data can be taken into account, allowing the learners to capture temporal dependency in multivariate time series data [27]. This time-delay embedding technique was first implemented in the TISEAN\(^1\) package for nonlinear time series analysis; however, such a package only processes univariate time series data.

In this work, an extension is made to one of their methods to handle time series data with more than one dimension. Specifically, for a given lag of $l$ time periods, an instance is represented with the most $l+1$ recent sets of attribute values. Hence, the size of the feature space would become $m \times (l+1)$. This time-delay embedding modification to expand the feature space to include historical data would allow the regressor to generate a regression model that also takes previous information into account, allowing temporal dependencies among multiple attributes to be modelled.

\(^1\)http://www.mpiik-dresden.mpg.de/ tisean
In this thesis, we propose to apply the time-delay embedding technique [27] which modifies the traditional feature space of machine learning algorithms so that history of data can be taken into account, allowing the learners to capture temporal dependency in multivariate time series data. This temporal-capable feature space allows a machine learning based regressor to learn the temporal relationship among observable attributes and the target latent attribute, while optimizing the prediction performance in a resource-efficient manner.

This section starts by describing our implementation framework, including data preprocessing, feature engineering (time-delay embedding), training and forecasting steps. The efficacy of the system is tested against the well-known baseline VARX model (VAR model with the target attribute as the exogenous variable). Then, we investigate the possibility of inferring the latent characteristics of an individual person using only his/her observable information. The best configuration of the proposed time series forecasting methods are trained with three types of information: observable information only, latent information only, and both. Then the performance is compared to make an empirical conclusion on the ability to predict emotion states from observable information.

Figure 3.1 outlines the methodology presented in this thesis. First, ground-truth data is acquired from participants. Since the raw data is not well-formed, a data preprocessing layer is applied to ensure that the data is in the proper multivariate time series format. The preprocessed data is then converted (i.e. featurized) into the format that machine learning algorithms can process. In this step, first, data points are generated, each of which represents a participant’s one day of data. Then, the time-delay embedding technique is applied to allow an instance of the training data to take temporal dependency into account. Once the data is featurized, a number of advanced machine learning based regression models are tested for their compatibility and ability to model the data (Objective 1). Then, the best forecaster from the Objective 1 is chosen to test the assumption that latent
attributes could effectively be predicted using only observable attributes. If this hypothesis is proven to be true, it would then be possible to implement a mental health monitoring system that only passively observes patients’ daily activities without interfering with their routines.

3.1 Data Acquisition

Ground truth validation data is collected from human participants via a series of questionnaires that can be conveniently accessible on smartphone devices. The questionnaires aim to quantify and monitor each participant’s daily emotional and physical states. Example questions are listed below:

- How many drinks did you have today?
- How many hours did you sleep last night?
- How many people did you interact with?
- How satisfied with life are you?
Do you enjoy the weather?

Table 3.1: Data associated with different attributes are collected at different frequencies.

<table>
<thead>
<tr>
<th>Attribute Type</th>
<th>Frequency</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Once</td>
<td>Pre-burst</td>
<td>Post-burst</td>
<td>Daily</td>
<td>Arbitrary</td>
</tr>
<tr>
<td>Observable</td>
<td>43</td>
<td>50</td>
<td>50</td>
<td>55</td>
<td>50</td>
</tr>
<tr>
<td>Unobservable</td>
<td>0</td>
<td>102</td>
<td>102</td>
<td>60</td>
<td>24</td>
</tr>
</tbody>
</table>

The questionnaires are chosen to quantify observable and latent attributes that have been shown in previous influential psychological literature to be effective and representative mental and physical conditions of human participants. A response to a questionnaire may result in one or more values of attributes being collected. For example, the response to the question How many hours of sleep did you get last night? will result in a numeric value associated with the attribute $SLPHRS_d$, which represents the number hours of last night’s sleep. Note that attributes with the suffix $d$ are daily attributes. Different questionnaires may be collected at different frequencies. For example, How old are you? may be asked once at the beginning of the data collection campaign, Did you exercise today? may be asked once a day, and How many friends did you just have a conversation with? may be asked arbitrarily every time the participant finishes a social interaction. Table 3.1 breaks down the numbers of attributes collected at different frequencies, classified by their observability. Listed below are the different frequencies where corresponding questionnaires are issued:

- **Once:** The corresponding questionnaires are asked only once at the beginning of the data collection campaign. Most of these inquiries pertain to background information such as gender, age, number of kids, employment status, etc.

- **Pre-burst:** The pre-burst questionnaires are distributed to each participant at the beginning of each burst of data collection. These questions aim to gauge the personality and physical health of each participant before a data collection session takes place.

- **Post-burst:** The post-burst questionnaires are the same as the pre-burst ones, and are used to measures the differences in personality and physical health of each participant during a data collection session.

- **Daily:** A participant is asked to respond to these daily questionnaires once a day. Most questions are designed to collect daily routines, and gauge mental and physical health.
Arbitrary: Participants are asked to respond to these arbitrary questionnaires after each social interaction event occurs. These questions are designed to keep track of certain properties of an interaction event (e.g., number of friends, how long, what time, purposes of the interaction, etc.). The intent of these social related questionnaires is to investigate whether social interaction could help to alleviate certain negative mental conditions.

Table 3.2 lists all the sets of questionnaires, each of which is intended to quantify certain characteristics of a participant. These sets of questionnaires can be categorized by their frequencies of response, namely once (background information), burst (pre/post burst visits), daily (daily information), and arbitrary (interaction/conversation information). Each set of questionnaires aims to quantify certain aspect of the participant. Each aspect is represented by a set of attributes. Response to each question will result in a value of a raw attribute. A composite attribute is calculated by combining certain relevant raw attributes. An example of a composite attribute includes the overall level of stress, which sums up the individual levels of stress from interpersonal tensions, work/school, home, finances, health/accident, other people’s events, and stress from being evaluated. Most of these questionnaires have been approved and used extensively for self-quantification of mental and physical characteristics (see references).

3.2 Data Preprocessing

Concretely, the data collected from a participant must be converted into the matrix format of multivariate time series data as illustrated in Equation 1.1 before further processing. From here on, it is assumed that a time unit is one day, so that the $i^{th}$ row of data represents the snapshot of data corresponding to the $i^{th}$ day. This data format would make sure that both traditional and newly-developed forecasting techniques can compatibly use the same data source, mitigating the bias (from data preparation) when comparing these algorithms.

Particularly, the forecasting models cannot apply on the data directly due to the following issues:

1. **Heterogeneous data types and ranges.** Different kinds of questions and options to respond were provided to the participants. An answer to a question can be binary (e.g., true/false), multiple choices, nominal values (e.g., day of week), non negative integer (e.g., number of friends), and ranged value (e.g., percentage). While most forecasting techniques only understand numeric attributes, each of these data types must first be converted to compatible, numeric values.

2. **Heterogeneous data frequencies.** As mentioned in Section 3.1, different attributes are collected at different frequencies (i.e., once, pre-burst, post-
Table 3.2: Sets of questionnaires, each aiming to quantify particular observable and latent characteristics of the participant, along with their references (if any).

<table>
<thead>
<tr>
<th>Group</th>
<th>ID</th>
<th>Name</th>
<th># Raw Attr</th>
<th># Comp Attr</th>
<th>Description</th>
<th>Ref(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DemoA</td>
<td>Demographics Visit 0</td>
<td>8</td>
<td>2</td>
<td>Background information such as gender, ethnic, employment status, marital status, income status, and zipcode.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DemoA</td>
<td>Demographics Visit 2-4</td>
<td>2</td>
<td>0</td>
<td>Information pertaining to stature and menopause regularity</td>
<td>[52, 60]</td>
</tr>
<tr>
<td></td>
<td>DemoC</td>
<td>Demographics Visit 4</td>
<td>21</td>
<td>1</td>
<td>Information about educational background, housing composition (especially children and their ages), medication (gastrointestinal and cardiovascular), height, and weight.</td>
<td>[56]</td>
</tr>
<tr>
<td></td>
<td>CESD</td>
<td>Center for Epidemiologic Studies Depression Scale</td>
<td>20</td>
<td>1</td>
<td>20 questions used to measure the overall level of depression.</td>
<td>[55]</td>
</tr>
<tr>
<td></td>
<td>BFI</td>
<td>Big Five Inventory</td>
<td>10</td>
<td>5</td>
<td>10 questions for self-measuring personality (Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness)</td>
<td>[57]</td>
</tr>
<tr>
<td></td>
<td>CONTROL</td>
<td>Control</td>
<td>9</td>
<td>1</td>
<td>Assessing sense of self-control and control over surrounding environments (relationship, finances, work, leisure, life).</td>
<td>[35]</td>
</tr>
<tr>
<td></td>
<td>CVI</td>
<td>Cardiovascular Health</td>
<td>8</td>
<td>2</td>
<td>Assessment of cardiovascular symptoms</td>
<td>[41]</td>
</tr>
<tr>
<td></td>
<td>GH</td>
<td>Gastrointestinal Health</td>
<td>12</td>
<td>2</td>
<td>Assessment of gastrointestinal disorders</td>
<td>[41]</td>
</tr>
<tr>
<td></td>
<td>META</td>
<td>Metacognition Questionnaire</td>
<td>15</td>
<td>1</td>
<td>15 questions for self-assessment of cognitive and neuropsychological impairment (distraction, focus, forgetfulness, etc.)</td>
<td>[5-6]</td>
</tr>
<tr>
<td></td>
<td>LES</td>
<td>Life Experiences</td>
<td>12</td>
<td>1</td>
<td>Assess of impacts of life changing events (i.e. relationship status, death of loved ones, illness, injuries, work status, finances, family members' well-being, pregnancy, incarceration, increase in level of arguments, eating/sleeping habits, living condition, stressful events.)</td>
<td>[28, 63]</td>
</tr>
<tr>
<td></td>
<td>SATIS</td>
<td>Overall Satisfaction with Life</td>
<td>6</td>
<td>1</td>
<td>Satisfaction with relationship, finances, health, work, leisure, and life.</td>
<td>[19, 53]</td>
</tr>
<tr>
<td></td>
<td>SF36</td>
<td>Short Form 36-item Health Survey Questionnaire</td>
<td>36</td>
<td>8</td>
<td>36 questions to self-assess mental and physical health</td>
<td>[46, 83]</td>
</tr>
<tr>
<td>Daily</td>
<td>LETQ</td>
<td>Leisure Time Questionnaire</td>
<td>3</td>
<td>1</td>
<td>Assessing level of today’s exercise (vigorous, moderate, mild)</td>
<td>[22]</td>
</tr>
<tr>
<td></td>
<td>SLEEP</td>
<td>Sleep</td>
<td>3</td>
<td>0</td>
<td>Assessing last night’s sleep quality.</td>
<td>[14]</td>
</tr>
<tr>
<td></td>
<td>WEATHER</td>
<td>Weather Enjoyment</td>
<td>1</td>
<td>0</td>
<td>Assessing weather enjoyment</td>
<td>[4]</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>Self Esteem</td>
<td>1</td>
<td>0</td>
<td>Level of self-esteem</td>
<td>[26]</td>
</tr>
<tr>
<td></td>
<td>CONTROL</td>
<td>Perceived Control</td>
<td>1</td>
<td>0</td>
<td>Measuring the level of overall control over things</td>
<td>[49]</td>
</tr>
<tr>
<td></td>
<td>SATIS</td>
<td>Satisfaction with Life and Health overall</td>
<td>2</td>
<td>0</td>
<td>Assessing levels of satisfaction with life and health.</td>
<td>[19]</td>
</tr>
<tr>
<td></td>
<td>SHAKE</td>
<td>State Shame and Guilt Scale</td>
<td>2</td>
<td>0</td>
<td>Assessing shame and guilt</td>
<td>[44]</td>
</tr>
<tr>
<td></td>
<td>STRESS</td>
<td>Perceived Stress Scale</td>
<td>11</td>
<td>2</td>
<td>Judging overall level of stress originated from environment such as interpersonal tensions, work/school, home, finances, health/accident, other people’s events, being evaluated, etc.</td>
<td>[16, 80]</td>
</tr>
<tr>
<td></td>
<td>FEELINGS</td>
<td>Feeling States</td>
<td>27</td>
<td>6</td>
<td>27 Dimensions of feelings, i.e. Enthusiastic, Calm, Nervous, Sluggish, Happy, Peaceful, Embarrassed, Sad, Alert, Satisfied, Upset, Bored, Proud, Relaxed, Depressed, Excited, Content, Fatigued, Tense, Disappointed, Ashamed, Relieved, Angry, Grateful, Conceited, Snobbish, and Successful</td>
<td>[11, 37, 45, 84]</td>
</tr>
<tr>
<td></td>
<td>PRIDE</td>
<td>State Pride Facets Scale</td>
<td>1</td>
<td>0</td>
<td>Today I felt successful</td>
<td>[72]</td>
</tr>
<tr>
<td></td>
<td>TU</td>
<td>Time Use</td>
<td>3</td>
<td>4</td>
<td>How much time did you spend on work/school, leisure, and other obligation.</td>
<td>[1, 38]</td>
</tr>
<tr>
<td></td>
<td>HB</td>
<td>Health Behaviors</td>
<td>9</td>
<td>2</td>
<td>Surveying health-related behaviors such as smoking, drinks, caffeine intakes, brushing, flossing, meals, and snacks.</td>
<td>[51]</td>
</tr>
<tr>
<td></td>
<td>EMOTION</td>
<td>State Emotions</td>
<td>7</td>
<td>0</td>
<td>Physical pain, attitude towards physical and emotional health, and expectation for tomorrow.</td>
<td></td>
</tr>
<tr>
<td>Arbitrary</td>
<td>Context variables</td>
<td>Arbitrary</td>
<td>15</td>
<td>0</td>
<td>Information about each interaction encountered, such as length, place, purpose, and interactants’ information (age, gender, relationship, acquaintanship).</td>
<td>[49]</td>
</tr>
<tr>
<td>Arbitrary</td>
<td>Interpersonal Grid</td>
<td>Arbitrary</td>
<td>4</td>
<td>0</td>
<td>Assessing interpersonal perceptions such as friendliness and dominance of self and others in a conversation.</td>
<td>[49]</td>
</tr>
<tr>
<td>Arbitrary</td>
<td>Affect Grid</td>
<td>Arbitrary</td>
<td>2</td>
<td>0</td>
<td>Assessing how you act (pleasant or unpleasant) and how you feel (sleepy or aroused).</td>
<td>[37]</td>
</tr>
<tr>
<td>Arbitrary</td>
<td>Emotion Regulation</td>
<td>Arbitrary</td>
<td>2</td>
<td>0</td>
<td>Assessing the abilities to suppress and reappraise emotions during the conversation.</td>
<td>[25]</td>
</tr>
<tr>
<td>Arbitrary</td>
<td>Interaction Reflections</td>
<td>Arbitrary</td>
<td>7</td>
<td>2</td>
<td>Assessing empathy (understanding others’ feelings) and the ability to measure cost/benefit from the conversation.</td>
<td>[25]</td>
</tr>
<tr>
<td>Arbitrary</td>
<td>State Emotions</td>
<td>Arbitrary</td>
<td>5</td>
<td>0</td>
<td>Assessing current emotions, i.e. anger, sadness, pride, shame, and happiness.</td>
<td></td>
</tr>
</tbody>
</table>

14
burst, daily, and arbitrary). While the value of each attribute must be available every day (since it is assumed that a time unit = one day), all attributes must be converted to daily attributes.

3. **Missing values.** Participants are not forced to answer the questions. Hence, it is inevitable that some or all questions on particular days may be left unanswered, resulting in missing values. While some forecasting techniques cannot handle missing values, they must be dealt with before further processing.

4. **Discontinuous data.** The data corresponding to a participant is divided into three bursts, resulting in three small chunks of multivariate time series data. While the forecasting models considered in this thesis do not have ensemble capability (where multiple learners, each of which may learn each chunk of data, then make an ensemble prediction), these chunks of data need to be concatenated to produce a smooth single time series data for each participant.

The subsequent subsections will describe how each case above is handled in this thesis.

### 3.2.1 Handling Different Data Types

If the value is already numeric (either bounded, or unbounded), then it remains the same. The unbounded number will not be an issue, since a normalization will be applied to the data before feeding it to the regression model. Hence, the values corresponding to each attribute would eventually be ranged to \([0, 1]\).

A value of each binary attribute (true or false) is replaced by either 0 or 1, based on the polarity. For example, **true** would be 1, and **false** would be 0.

A multiple choice or nominal attribute is first converted to a sequence of binary attributes, where the above solution for binary attributes can be applied. For example, a value of attribute \(A\) could be one of the days of week. \(A\) is first split into 7 sub-attributes: \(A\).sun, \(A\).mon, \(A\).tue, \(A\).wed, \(A\). thu, \(A\).fri, and \(A\).sat, each of which is a binary attribute. Hence, if the original value is **Monday**, then the sub-attribute \(A\).mon would be 1, while the other sub-attributes become 0.

### 3.2.2 Handling Different Data Frequencies

Since some attributes may be collected less or more than once a day, these attributes must be normalized so that their values can be available on a daily basis. For attributes whose values are collected less than daily (i.e. once, pre-burst, and post-burst), their values are replicated on subsequent days, where applicable. For attributes whose values are collected more than once a day, their values corresponding to the same day are aggregated (by summation if numeric, and median
if nominal). These aggregated values are then used to represent daily values of the corresponding non-daily attributes.

### 3.2.3 Handling Missing Values

Multiple schemes have been used to deal with missing attribute values including using the default values (e.g., 0) and completely discarding the instances containing missing values [62]. Here, the missing values are cubic spline interpolated using the data available from the same participant. If there is not enough data to interpolate the missing values, then they are set to default values.

### 3.2.4 Handling Discontinuous Data

For each participant, the three bursts of data are concatenated while preserving the sequence of the days of week. For example, if the first burst of data collection ends on Friday, while the beginning day of the second burst is Wednesday, then the last three days of the data of the first burst (i.e., Wednesday, Thursday, and Friday) are discarded. The sequence of days of week is preserved because, in some cases, the learners keep track of seasonality. For example, if the seasonality is weekly (7 days), some learners are able to predict that the set of situations that happen on a Friday would likely happen again on the next Friday.

### 3.3 Feature Space Modification

The proposed machine learning based multivariate time series forecasting methodology in this work relies on the use of base machine learning regressors to train on the history of multi-variate time series data. However, traditional machine learners treat an instance (daily data) independently, which makes sense since most types of data points in traditional machine learning literature are assumed to be independent (e.g., images, documents, emails, etc.). However, in this work, each data point represents a snapshot of attribute values of a participant collected on each day, which can be dependent on the values of prior days, in terms of both seasonality and activity decay/growth. An example of seasonality would be that a participant may habitually drink heavily on Fridays, moderately on weekends, and not at all on the other week days. Hence, if such a participant happens to drink heavily on a work day, then the learner should be skeptical about possible mental anomalies. An example of activity decay would be the level of fever which can reduce in its magnitude every day. An example of activity growth would be the accumulative stress from final exams, which tends to gradually increase as the final exam week draws near.

In order to take the seasonality and activity decay/growth into account, an instance of data must incorporate the temporal relationship between the data
point at a given day and the data points of the previous days. The number of previous days that a data point is relating to is referred to as lag. Equivalently, the \(i^{th}\) data point with lag \(l\) is the data point representing the set of attributes whose values are collected at the \(i^{th}\) day, along with the previous \(l\) days. A data point with lag 0 only represents current (today’s) values of attributes.

Here, the time-delay embedding algorithm is applied to the feature space by expanding the slots for previous data associated with the participant \(p_k\). Mathematically, let the instance \(p_k(i)\) represents the data snapshot of the \(i^{th}\) day:

\[
p_k(i) = [a^{(1)}_i, a^{(2)}_i, a^{(3)}_i, \ldots, a^{(m)}_i] \quad (3.1)
\]

Then the time-delay embedded version of such an instance with lag = \(l\), \(p_k(i, l)\), is defined as:

\[
p_k(i, l) = \begin{bmatrix}
  a^{(1)}_i & a^{(2)}_i & a^{(3)}_i & \ldots & a^{(m)}_i \\
  a^{(i-1)}_1 & a^{(i-1)}_2 & a^{(i-1)}_3 & \ldots & a^{(i-1)}_m \\
  a^{(i-2)}_1 & a^{(i-2)}_2 & a^{(i-2)}_3 & \ldots & a^{(i-2)}_m \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  a^{(i-l)}_1 & a^{(i-l)}_2 & a^{(i-l)}_3 & \ldots & a^{(i-l)}_m
\end{bmatrix}
\]

Basically, for each attribute \(a^{(i)}_j\), the feature space also includes its previous \(l\) values. Once the feature space is modified according to the rules above, a traditional machine learning based regressor can then learn and predict using the conventional machine learning regression method.

### 3.4 Objective 1: Model Section

In this work, we present a machine learning based forecasting methodology for multivariate time series data. We claim that such methodology is built upon machine learning algorithms, some of which are known to handle high dimensionality and non-linear relationships quite effectively [36]. In this thesis, such claims are tested by comparing the forecasting efficacy of our proposed methods with the traditional VARX model.

Many time series forecasting techniques have been proposed for a wide range of forecasting problems. In this thesis, 10 time series forecasting models are considered, including a variant of traditional vector autoregression (VARX) and multiple machine learning based regression techniques from different families such as function based, tree based, and lazy learning based methods. These models are listed in Table 3.3, along with their references.
Table 3.3: List of forecasting models considered, along with their references.

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Model Name</th>
<th>Algorithm Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>VARX</td>
<td>Vector Auto-regression [74]</td>
<td>Regression</td>
</tr>
<tr>
<td>RF</td>
<td>Regression by Discretization using Random Forest [12,71]</td>
<td>Machine Learning: Tree Based</td>
</tr>
<tr>
<td>M5P</td>
<td>M5 Model Tree with Continuous Class Learner [54,82]</td>
<td>Machine Learning: Tree Based</td>
</tr>
</tbody>
</table>

The methodology in this objective is carried out in two stages: partial and comprehensive. In the partial stage, all forecasting models listed in Table 3.3 are trained with partial, dimension reduced data, and tested against each other. Participants with incomplete (missing) information are disregarded, and the remaining data is projected onto a lower dimension space (i.e. 10 dimensions) using the Principle Component Analysis algorithm [30]. Note that preprocessing the data with dimensional reduction techniques could eliminate necessary information that could have been captured by the time series models; however, in this stage, the dimension of the data is reduced to allow fair comparison between the VARX model and the machine learning based models. Only complete data is chosen to ensure that the models are not tested on their ability to handle missing values. This dimension reduction is carried out because the VARX technique cannot handle data with large dimension\(^1\). The goal of the partial stage is to seek for the best forecasting models that work on the dataset used in this thesis.

The comprehensive stage selects top forecasting models from the partial stage, then run them on full data, with missing values and complete dimension. The goal of this stage is to find the best model for each target attribute, and for analysis of the results in the Objective 2.

---
\(^1\) In our experiments, Matlab would hang or throw error messages when VARX is used to model more than 10 attributes of data.
3.5 Objective 2: Predicting Latent Attributes

This thesis investigates the possibility to train the forecaster with only observable information, in order to predict the latent attributes (e.g., feelings, mental statuses, etc.). These latent attributes are difficult to quantify merely just externally observing; hence, the ability to infer and forecast them could prove to be valuable in multiple computational psychology related applications. The best forecaster from the Objective 1 is selected for the analysis in this objective. We prepare three types of training data: observable only (O), latent only (U), and both (OU). The forecaster is trained with each of these source types, then the forecasting performance of the select latent attributes is compared. We first investigate the impact of different data sources on the forecasting performance. Then, we vary different values of lags to study what would be the appropriate amount of the history of data that the learner should keep track of.
Case Study

For each participant $p_k$, and each target attribute $a_t$, a forecasting model is trained with the history of data of $p_k$ for the lag $l$ days to predict the value of $a_t$ on the next $h$ days in the future. The history of $a_t$ is treated as an external variable. That is, it is not included in the training data and its historical data is not used. Target attributes are made external variables because it is an assumption in this work that these attributes are not easily quantified through simple external observation. Hence, the ability to estimate these attributes accurately would be more valuable than that of the observable attributes.

4.1 Ground Truth Validation Data

Figure 4.1: Distribution of the Positive Effect ($PosEffect.D$) attributes of the 150 participants. The Y axis marks the level (0 to 100), and the X axis represents ages in years of the participants. Each line is corresponding to a participant. Each dot is represents a daily positive effect level.

A data collection campaign was conducted among 150 participants from the Pennsylvania State University Community, ages range from 18 - 90 years (mean
51% of the participants are women, and 49% are men. 91% are Caucasian, 4% African American, 1% Asian American, 2% Other, 2% mixed (Asian, Hispanic/Latino, or American Indian + Caucasian). 93% of the participants determined themselves as heterosexual, 6% Bisexual/Gay/Lesbian, and 1% declined to reveal. Each participant has an average number of children of 1.5 (standard deviation 1.41, min 0, max 6).

Figure 4.2: Illustration of the three bursts during which the data was collected.

The data was collected in three bursts, each of which lasted for 21 days. There was a 4.5 month break between the first and second bursts, and 3.5 month break between the second and third bursts. Each participant is required to participate in a 3 hour pre-burst assessment and training session which gives an overview of the project and provides instructions to install and use the mobile app that is used for collecting the participants’ responses. Furthermore, each participant is also required to participate in the 30 minute post-burst assessment session for verification purposes. Figure 4.2 illustrates the three bursts of data collection process. For further details about data acquisition procedure, please refer to the iSAHIB project.

As an example of the collected attributes, Figure 4.1 plots the ranges of the Positive Effect ($PosEffect_d$) attribute against ages of all the 150 participants. Each line is corresponding to a participant. There are roughly 60 dots in each line, each of which represents the positive effect level measured in a day. Each attribute leads to a new dimension in the time series data for each participant.

### 4.2 Selected Target Attributes

Table 4.1 lists the 6 target attributes used as test cases in this research. These attributes are all latent attributes. All of these attributes (except Physical Health) have been used in previous studies (the references are provided at each attribute’s description). These 6 attributes are chosen as representative target attributes, that represent different aspects of mental conditions of a participant. Note that

1. http://studiolab.psu.edu/projects/isahib
Table 4.1: List of select target attributes along with their references. Refer to Table 3.2 for the comprehensive description of each attribute.

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONTROL_d</td>
<td>Perceived Control. (Did you have control over the things that happened to you today?) [39]</td>
</tr>
<tr>
<td>NegAffect_d</td>
<td>Net effect of negative emotional attributes such as nervousness, disappointment, boredom, etc. [11,37,45,84]</td>
</tr>
<tr>
<td>PHYHEALTH_d</td>
<td>Perception of physical health.</td>
</tr>
<tr>
<td>PosAffect_d</td>
<td>Net effect of positive emotional attributes such as pride, calm, happiness, etc. [11,37,45,84]</td>
</tr>
<tr>
<td>SATHEALTH_d</td>
<td>Satisfaction with health. [19,53]</td>
</tr>
<tr>
<td>SATLIFE_d</td>
<td>Satisfaction with life. [19,53]</td>
</tr>
</tbody>
</table>

if a composite attribute is chosen as the target attribute, then all the sub (raw) attributes used to produce it are not taken into the feature space to avoid the causality and dependency biases.

4.3 Implementation

The experiments in this thesis use the Matlab’s implementation of the VARX model\(^2\), and Weka’s implementation of all other base machine learning classification and regression algorithms\(^3\). Listed below are the major implementations during this thesis (using Java as the main programming language):

1. Data import, cleaning, pre-processing, and conversion to the compatible multivariate time series data structure.

2. The time-delay embedding algorithm which flattens multi-variate time series data into single dimension time series data, with control parameter lag \(l\).

3. A wrapper that enables each base learner to incrementally learn new instances without having to retrain the whole model from scratch. This is useful when conducting the leave-one-out sliding evaluation protocol in which the current validation data is fed back to forecaster’s training data to make the next prediction, until no ground truth data is available.

4. Experimental framework, including batch commands, result logging, and computation of evaluation statistics.

\(^2\)http://www.mathworks.com/help/econ/vgxvarx.html

\(^3\)http://www.cs.waikato.ac.nz/ml/weka/
Note that, though Weka also offers a forecasting package, such a package would not facilitate our study for the following reasons:

1. If the user wants to model the dependency among multiple attributes, these attributes must be set as target attributes. The tool would then generate a multivariate model for each target attribute. In our setting that involves modelling hundreds of attributes, it would be computationally expensive to generate a model for each of them.

2. The tool does not offer leave-one-out sliding evaluation protocol. Though the user could set apart a certain portion of the training data for testing, the tool does not feed some of already-tested data to the training data. In our evaluations, we would like to feed the already-tested data back to the training data to predict further not-yet-tested values to see whether the error could be reduced by having more training data.

For these reasons, it would be more efficient to implement our own multivariate time series framework for flexibility and future development purposes. The source code will also be available to others for research purposes (upon request).

4.4 Forecasting Evaluation Protocol

Leave-one-out sliding evaluation protocol is carried to validate the forecasting models. Such an evaluation protocol is widely used to evaluate time series forecasting models [32]. For each participant $p_k$, each target attribute $a_t$, lag $l$, and a given day $i$, the forecasting model learns the history of the data from day $(i - l)$ to $(i)$, then makes the prediction for the target value of the next $h$ days. In the experiment, the predictions are made for each of the most recent 14 values in the time series (roughly the latter half of the data in the 3rd burst), then the statistics of the predictions corresponding to all the individual participants are averaged.

4.5 Forecasting Evaluation Metrics

Five metrics are used to compare and quantify the forecastability of each model, including directional accuracy (DAC), mean average error (MAE), mean average percentage error (MAPE), mean square error (MSE), root mean square error (RMSE), and model training time (TrainTime). Such metrics have been successfully used to measure forecastability in multiple forecasting related works [76, 77].
4.6 Objective 1: Model Selection

This section reports the results from comparing multiple forecasting models in order to select the best one for further experiments.

4.6.1 Stage 1: Partial Comparison

The analysis of three chosen attributes are shown (in Table 4.2). It is apparent that the performance of machine learning based methods (especially Random Forest) outperform the baseline VAR model on all the three chosen attributes. It is interesting to see that the traditional VARX method that has been widely used for multivariate time series forecasting performs the worst, compared to the other machine learning based algorithms. This could be not only because the data has so many dimensions that the VARX model can effectively handle, but also because the relationship between the attributes may not be linear.

4.6.2 Stage 2: Comprehensive Comparison

In this section, the top regressors are chosen to test on the full data, without dimensionality reduction.

Table 4.3 lists the forecasting performance of top forecasting from Stage 1, on the six select attributes. These models are trained and tested on the full data without dimension reduction. Red/bold figures are the best (lowest error, highest directional accuracy) evaluation scores among all the forecasting models. It is evident that Random Forest forecasters yield the lowest error (as measured by MAE, MAPE, MSE, and RMSE). This finding is also consistent with the Stage 1, where all the forecasters are trained/tested on dimension reduced, partial data. Random Forest has been shown to be successful in many classification tasks such as [75, 78, 79], due to its ability to deal with unbalanced data, avoid over-fitting, and automatically select useful features. In terms of directional accuracy (DAC), Random Forest performs the best only for the attribute $SATHEALTH_d$, while Radial Basis Function Networks for Regression (RBFN) outperforms the other on the remaining five target attributes. However, Random Forest stands the second best in most of such attributes. Also, the DAC differences between the RBFN and the Random Forest are only marginal. Hence, in terms of forecasting performance, we believe the forecasters implementing the Random Forest algorithm are the most suitable ones for our dataset.

It may be the case that good performance comes at the cost of learning time. Though the training times of Random Forest are not as large as those of Linear Regression (LR) and Multi-layer Perceptron for Regression (MPR), due to not having built-in feature selection mechanism, they are still quite computationally expensive compared to other relatively good models, such as RBFN and M5P. This
Table 4.2: Comparison of the forecasting results of the three sample target attributes on the select forecasting models with lag $l = 3$ and forecasting horizon $h = 1$. The models are trained with partial training data (i.e. using dimension-reduced data with no missing values).

<table>
<thead>
<tr>
<th>Target Attribute</th>
<th>Reg. Model</th>
<th>DAC</th>
<th>MAE</th>
<th>MAPE (%)</th>
<th>MSE</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>NegAffect</strong></td>
<td>VAR</td>
<td>0.5257</td>
<td>16.5378</td>
<td>28.4516</td>
<td>459.9315</td>
<td>18.2033</td>
</tr>
<tr>
<td></td>
<td>SVMR</td>
<td>0.6327</td>
<td>6.3261</td>
<td><strong>15.0258</strong></td>
<td>84.4304</td>
<td>7.8731</td>
</tr>
<tr>
<td></td>
<td>SLR</td>
<td>0.6749</td>
<td>6.3707</td>
<td>16.2492</td>
<td>79.7906</td>
<td>7.7339</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.6893</td>
<td>6.0867</td>
<td>66.9891</td>
<td><strong>72.3964</strong></td>
<td><strong>7.4043</strong></td>
</tr>
<tr>
<td></td>
<td>RBFN</td>
<td>0.6862</td>
<td>6.6261</td>
<td>17.3026</td>
<td>87.6746</td>
<td>8.0549</td>
</tr>
<tr>
<td></td>
<td>MPR</td>
<td>0.6348</td>
<td>8.1201</td>
<td>18.9394</td>
<td>137.1940</td>
<td>10.3114</td>
</tr>
<tr>
<td></td>
<td>M5P</td>
<td>0.6646</td>
<td>6.3734</td>
<td>15.1783</td>
<td>80.4128</td>
<td>7.7991</td>
</tr>
<tr>
<td></td>
<td>LWL</td>
<td>0.6759</td>
<td>6.5596</td>
<td>15.5309</td>
<td>88.7346</td>
<td>8.1442</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>0.6512</td>
<td>6.4480</td>
<td>15.2470</td>
<td>80.8905</td>
<td>7.8626</td>
</tr>
<tr>
<td></td>
<td>Ibk</td>
<td>0.5874</td>
<td>7.5983</td>
<td>16.5657</td>
<td>118.1801</td>
<td>9.4736</td>
</tr>
<tr>
<td></td>
<td>GPR</td>
<td>0.6564</td>
<td>6.1359</td>
<td>15.4226</td>
<td>74.1160</td>
<td>7.4966</td>
</tr>
<tr>
<td><strong>PosAffect</strong></td>
<td>VAR</td>
<td>0.5535</td>
<td>26.3600</td>
<td>71.0436</td>
<td>1095.0173</td>
<td>30.2500</td>
</tr>
<tr>
<td></td>
<td>SVMR</td>
<td>0.6461</td>
<td>6.5107</td>
<td>15.4641</td>
<td>85.3383</td>
<td>8.2210</td>
</tr>
<tr>
<td></td>
<td>SLR</td>
<td>0.6667</td>
<td>6.4836</td>
<td>16.5372</td>
<td>79.3142</td>
<td>8.0352</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.6842</td>
<td>5.8792</td>
<td><strong>14.7563</strong></td>
<td><strong>66.5746</strong></td>
<td><strong>7.3716</strong></td>
</tr>
<tr>
<td></td>
<td>RBFN</td>
<td>0.6759</td>
<td>6.6577</td>
<td>17.3851</td>
<td>83.5140</td>
<td>8.1976</td>
</tr>
<tr>
<td></td>
<td>MPR</td>
<td>0.6235</td>
<td>8.4115</td>
<td>19.6191</td>
<td>144.4885</td>
<td>10.6435</td>
</tr>
<tr>
<td></td>
<td>M5P</td>
<td>0.6595</td>
<td>6.3026</td>
<td>15.0097</td>
<td>74.6272</td>
<td>7.8280</td>
</tr>
<tr>
<td></td>
<td>LWL</td>
<td>0.6584</td>
<td>6.6242</td>
<td>15.6839</td>
<td>84.2575</td>
<td>8.1982</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>0.6533</td>
<td>6.3524</td>
<td>15.0209</td>
<td>74.7017</td>
<td>7.8534</td>
</tr>
<tr>
<td></td>
<td>Ibk</td>
<td>0.5710</td>
<td>7.6001</td>
<td>16.5697</td>
<td>117.0639</td>
<td>9.5517</td>
</tr>
<tr>
<td></td>
<td>GPR</td>
<td>0.6615</td>
<td>6.1047</td>
<td>15.3441</td>
<td>72.4103</td>
<td>7.6512</td>
</tr>
<tr>
<td><strong>Control</strong></td>
<td>VAR</td>
<td>0.5288</td>
<td>29.4296</td>
<td>79.3165</td>
<td>1428.1814</td>
<td>33.4424</td>
</tr>
<tr>
<td></td>
<td>SVMR</td>
<td>0.6759</td>
<td>8.6540</td>
<td>20.5549</td>
<td>162.4305</td>
<td>10.8847</td>
</tr>
<tr>
<td></td>
<td>SLR</td>
<td>0.6759</td>
<td>8.4926</td>
<td>21.6613</td>
<td>149.1373</td>
<td>10.5181</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.6739</td>
<td><strong>8.0625</strong></td>
<td>20.8032</td>
<td><strong>130.8079</strong></td>
<td><strong>9.9277</strong></td>
</tr>
<tr>
<td></td>
<td>RBFN</td>
<td>0.6656</td>
<td>8.8722</td>
<td>23.1677</td>
<td>156.0613</td>
<td>10.8053</td>
</tr>
<tr>
<td></td>
<td>MPR</td>
<td>0.6183</td>
<td>11.1605</td>
<td>26.0310</td>
<td>285.8279</td>
<td>14.7689</td>
</tr>
<tr>
<td></td>
<td>M5P</td>
<td>0.6718</td>
<td>8.5195</td>
<td><strong>20.2893</strong></td>
<td>150.1819</td>
<td>10.6231</td>
</tr>
<tr>
<td></td>
<td>LWL</td>
<td>0.6708</td>
<td>9.0107</td>
<td>21.3344</td>
<td>184.7712</td>
<td>11.4806</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>0.6749</td>
<td>8.5829</td>
<td>20.2951</td>
<td>152.7628</td>
<td>10.6900</td>
</tr>
<tr>
<td></td>
<td>Ibk</td>
<td>0.5885</td>
<td>10.0613</td>
<td>21.9355</td>
<td>216.7923</td>
<td>12.6711</td>
</tr>
<tr>
<td></td>
<td>GPR</td>
<td><strong>0.6770</strong></td>
<td>8.4783</td>
<td>21.3100</td>
<td><strong>144.5252</strong></td>
<td><strong>10.4939</strong></td>
</tr>
</tbody>
</table>

Computational resource consumption due to internal configuration of the Random Forest, which builds 300 atomic decision trees to make ensemble decisions. However, since the training process can be done off-line, we will stick with Random
Table 4.3: Average forecasting results of all the 150 participants using full data ($L = 7$ and $H = 1$) of each some forecasting models on the select six attributes.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Model</th>
<th>DAC</th>
<th>MAE</th>
<th>MAPE</th>
<th>MSE</th>
<th>RMSE</th>
<th>Train Time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LR</td>
<td>0.6132</td>
<td>10.86</td>
<td>25.73</td>
<td>251.82</td>
<td>14.10</td>
<td>2272.28</td>
</tr>
<tr>
<td></td>
<td>MPR</td>
<td>0.5936</td>
<td>11.24</td>
<td>26.36</td>
<td>268.82</td>
<td>14.64</td>
<td>2017.06</td>
</tr>
<tr>
<td></td>
<td>GPR</td>
<td>0.6368</td>
<td>10.29</td>
<td>23.31</td>
<td>242.02</td>
<td>12.80</td>
<td>27.45</td>
</tr>
<tr>
<td></td>
<td>MSP</td>
<td>0.6584</td>
<td>9.27</td>
<td>22.80</td>
<td>178.22</td>
<td>11.82</td>
<td>374.15</td>
</tr>
<tr>
<td></td>
<td>RBFN</td>
<td>0.6749</td>
<td>8.47</td>
<td>22.71</td>
<td>145.26</td>
<td>10.40</td>
<td>172.64</td>
</tr>
<tr>
<td></td>
<td>J48</td>
<td>0.6492</td>
<td>10.09</td>
<td>27.42</td>
<td>211.61</td>
<td>12.72</td>
<td>127.64</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.6872</td>
<td>8.12</td>
<td>21.56</td>
<td>134.05</td>
<td>10.04</td>
<td>458.31</td>
</tr>
<tr>
<td></td>
<td>SMVR</td>
<td>0.6348</td>
<td>10.40</td>
<td>23.49</td>
<td>234.70</td>
<td>12.98</td>
<td>34.27</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>LR</td>
<td>0.5730</td>
<td>14.27</td>
<td>40.62</td>
<td>401.74</td>
<td>17.98</td>
<td>4355.01</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>MPR</td>
<td>0.5710</td>
<td>12.96</td>
<td>38.82</td>
<td>340.00</td>
<td>16.39</td>
<td>2832.36</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>GPR</td>
<td>0.5905</td>
<td>11.66</td>
<td>36.41</td>
<td>308.97</td>
<td>14.60</td>
<td>26.70</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>M5P</td>
<td>0.6101</td>
<td>10.77</td>
<td>32.32</td>
<td>249.47</td>
<td>13.75</td>
<td>127.64</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>RBFN</td>
<td>0.6492</td>
<td>10.44</td>
<td>35.83</td>
<td>261.73</td>
<td>12.81</td>
<td>38.87</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>J48</td>
<td>0.6142</td>
<td>10.67</td>
<td>34.57</td>
<td>241.46</td>
<td>13.25</td>
<td>103.60</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>RF</td>
<td>0.6440</td>
<td>9.11</td>
<td>21.56</td>
<td>134.05</td>
<td>11.82</td>
<td>389.56</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>SMVR</td>
<td>0.5957</td>
<td>11.59</td>
<td>36.36</td>
<td>300.03</td>
<td>14.55</td>
<td>33.53</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>LR</td>
<td>0.6060</td>
<td>11.14</td>
<td>28.77</td>
<td>258.70</td>
<td>14.12</td>
<td>2873.44</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>MPR</td>
<td>0.6080</td>
<td>11.35</td>
<td>26.62</td>
<td>273.51</td>
<td>14.17</td>
<td>2816.51</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>GPR</td>
<td>0.5916</td>
<td>10.15</td>
<td>23.53</td>
<td>260.39</td>
<td>12.55</td>
<td>26.80</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>M5P</td>
<td>0.6101</td>
<td>9.51</td>
<td>23.33</td>
<td>200.62</td>
<td>11.85</td>
<td>26.70</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>RBFN</td>
<td>0.6153</td>
<td>9.60</td>
<td>24.28</td>
<td>184.27</td>
<td>11.24</td>
<td>40.54</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>J48</td>
<td>0.6173</td>
<td>10.42</td>
<td>25.90</td>
<td>226.56</td>
<td>12.84</td>
<td>121.99</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>RF</td>
<td>0.6285</td>
<td>8.37</td>
<td>20.94</td>
<td>140.32</td>
<td>10.02</td>
<td>451.94</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>SMVR</td>
<td>0.5936</td>
<td>10.21</td>
<td>23.67</td>
<td>262.68</td>
<td>12.63</td>
<td>33.87</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>LR</td>
<td>0.6276</td>
<td>8.22</td>
<td>83.94</td>
<td>139.99</td>
<td>10.40</td>
<td>1840.27</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>MPR</td>
<td>0.6080</td>
<td>8.32</td>
<td>74.46</td>
<td>138.72</td>
<td>10.51</td>
<td>1862.34</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>GPR</td>
<td>0.6481</td>
<td>7.72</td>
<td>74.31</td>
<td>176.24</td>
<td>9.74</td>
<td>24.26</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>M5P</td>
<td>0.6502</td>
<td>6.96</td>
<td>70.73</td>
<td>98.93</td>
<td>8.67</td>
<td>340.22</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>RBFN</td>
<td>0.6965</td>
<td>6.38</td>
<td>68.68</td>
<td>84.84</td>
<td>7.79</td>
<td>40.31</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>J48</td>
<td>0.6615</td>
<td>7.30</td>
<td>68.39</td>
<td>108.83</td>
<td>9.07</td>
<td>135.65</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>RF</td>
<td>0.6883</td>
<td>6.06</td>
<td>65.63</td>
<td>72.38</td>
<td>7.35</td>
<td>504.95</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>SMVR</td>
<td>0.6440</td>
<td>7.81</td>
<td>75.87</td>
<td>182.38</td>
<td>9.85</td>
<td>32.46</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>LR</td>
<td>0.6039</td>
<td>8.95</td>
<td>22.00</td>
<td>169.22</td>
<td>11.42</td>
<td>7912.66</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>MPR</td>
<td>0.5854</td>
<td>8.88</td>
<td>21.21</td>
<td>148.73</td>
<td>11.05</td>
<td>3642.25</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>GPR</td>
<td>0.6029</td>
<td>8.02</td>
<td>18.41</td>
<td>166.55</td>
<td>9.96</td>
<td>27.98</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>M5P</td>
<td>0.6451</td>
<td>7.29</td>
<td>17.19</td>
<td>106.12</td>
<td>9.17</td>
<td>383.06</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>RBFN</td>
<td>0.6944</td>
<td>6.49</td>
<td>16.62</td>
<td>79.38</td>
<td>7.39</td>
<td>39.32</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>J48</td>
<td>0.6183</td>
<td>7.68</td>
<td>19.31</td>
<td>111.20</td>
<td>9.52</td>
<td>128.10</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>RF</td>
<td>0.6883</td>
<td>5.87</td>
<td>14.83</td>
<td>66.12</td>
<td>7.36</td>
<td>509.24</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>SMVR</td>
<td>0.6008</td>
<td>8.09</td>
<td>18.56</td>
<td>169.12</td>
<td>10.04</td>
<td>36.81</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>LR</td>
<td>0.6163</td>
<td>10.02</td>
<td>26.21</td>
<td>221.13</td>
<td>12.72</td>
<td>2350.12</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>MPR</td>
<td>0.5751</td>
<td>10.93</td>
<td>25.48</td>
<td>243.56</td>
<td>13.69</td>
<td>2095.35</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>GPR</td>
<td>0.6327</td>
<td>9.29</td>
<td>23.27</td>
<td>222.05</td>
<td>11.73</td>
<td>28.49</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>M5P</td>
<td>0.6502</td>
<td>8.51</td>
<td>26.04</td>
<td>156.80</td>
<td>10.83</td>
<td>377.62</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>RBFN</td>
<td>0.6842</td>
<td>7.91</td>
<td>24.46</td>
<td>129.36</td>
<td>9.73</td>
<td>39.49</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>J48</td>
<td>0.6461</td>
<td>9.12</td>
<td>25.87</td>
<td>173.91</td>
<td>11.46</td>
<td>139.99</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>RF</td>
<td>0.6813</td>
<td>7.26</td>
<td>22.08</td>
<td>113.20</td>
<td>9.14</td>
<td>466.79</td>
</tr>
<tr>
<td>CRTL_d</td>
<td>SMVR</td>
<td>0.6307</td>
<td>9.46</td>
<td>23.49</td>
<td>234.20</td>
<td>11.90</td>
<td>35.24</td>
</tr>
</tbody>
</table>

Forest for the analysis in the Objective 2. In future works, another implement-
tation of Random Forest, *Fast Random Forest*\(^4\), which was claimed to improve upon the current implementation used in this thesis (speed is one of the major improvements) could be explored.

Figure 4.3: (Left) Comparison of example forecasting results by RF with different horizons (i.e. \(h = 0, 1, 2, 3\)) against the actual values of the attribute \(SATLIFE_d\) of a participant. (Right) Comparison of absolute error of each horizon (i.e. \(h = 0, 1, 2, 3\)).

Figure 4.3 (Left) shows sample forecasting results of the actual values of a participant on the attribute \(SATLIFE_d\). The forecasting model used here is Random Forest, trained with 3 days of historical data (\(l = 3\)). The prediction horizons vary from 0 to 3. Note that \(h = 0\) means the forecasting model is predicting today’s value of the target attribute. This particular example illustrates that the prediction accuracy decreases as the model predicts the value further ahead in the future. Figure 4.3 (Right) shows the absolute errors calculated from the predictions in the left figure. It is interesting to note that the absolute error decreases as the model predicts more recent values. This is because as the forecaster proceeds to predict the value in the next period, the current values are fed back to the training data, resulting in more historical data to learn from.

### 4.7 Objective 2: Predicting Latent Attributes

In this objective, we investigate whether it is possible to predict the latent attributes using only the observable information. First the impact of different training data sources (i.e. observable only, latent only, and both) is investigated. Then we study the effect of different lags on the forecasting performance.

\(^4\)https://code.google.com/p/fast-random-forest/
4.7.1 Impact of Different Sources of Information

Figure 4.4: Comparison of average mean absolute error (MAE) produced by Random Forest forecasters trained with different information sources (i.e. OU = both observable and latent information, O = only observable information, U = only latent information) at each horizon (days ahead of prediction). Each prediction is an average of prediction using different lags (i.e. lags = 1, 3, 5, and 7). Each attribute value has a range of [0, 100].

The ability to predict and forecast latent attributes using the information from
only observable attributes could prove crucial to multiple applications in mental and emotion detection. For each participant, a Random Forest forecaster is trained with each of the three data source modes: observable only (O), latent only (U), and both (OU). For each source mode, the lags are varied between 1, 3, 5, and 7. The results from all the participants and lags are averaged.

Figure 4.4 plots the mean absolute error (MAE) of the prediction at different horizons on the six selected latent attributes. It is a single consensus that the forecasters trained with only observable information (blue-circle) perform the best, as they achieve the lowest absolute error at every horizon for all the select attributes. Note that since each select latent attribute can take a value from \([0 - 100]\), the magnitude of error can be thought of as percentage error (in the absolute sense). The magnitudes of error vary across different latent attributes. For instance, the forecasts for the attribute \(PHYHEALTH_d\) have the absolute error fluctuating around \([8.5 - 9]\), while those of \(NegEffect_d\) are fluctuating around \([6.0 - 6.1]\).

Regardless, the errors are considered small and acceptable, suggesting that observable daily routine and behavior information can be good predictors to measure mental and emotion conditions. Surprisingly, the forecasters trained only with latent information (magenta-square lines) perform the worst. This suggests that the emotional attributes could be the causal effects of the physical ones, making the relationship among latent attributes rather loose or non-existent. It is interesting to note that the error magnitudes of the prediction of the forecasters trained with only latent information significantly and constantly decrease as the horizon increases, despite the intuition that further prediction should be less accurate. Regardless of which, the magnitudes of errors of these latent-only forecasters are quite large, compared to the observable-only ones. The forecasters trained with both observable and latent information perform somewhere in between. This is reasonable because, while the observable information is proven most useful, adding the latent information could taint the learned model, impeding the forecasting accuracy. Though Random Forest has built-in feature selection mechanism, the effect of the addition of this less useful information may not be completely eliminated.

Figure 4.5 shows the average directional accuracy of the forecasting of the select six latent attributes. Recall that directional accuracy quantifies the accuracy of the predicted polarity change (i.e., greater or less than the current value) of a particular attribute. Unlike the performance in terms of error magnitudes, where the observable-only forecasters all outperform other forecasters, the directional accuracy results indicate a mixed conclusion. The observable-only forecasters perform better than others on the attribute \(SATHEALTH_d\), and on-par with the combined forecasters on the attributes \(CONTROL_d, PHYHEALTH_d, PosAffect_d,\) and \(SATLIFE_d\). The directional accuracy of the observable only is partially worst on the attribute \(NegEffect_d\).

In this section, we train the Random Forest forecaster with three sources of information: observable only, latent only, and both. The results in terms of error...
Figure 4.5: Comparison of directional accuracy (DAC), range between [0, 1], produced by Random Forest forecaster trained with different information sources (i.e. OU = both observable and latent information, O = only observable information, U = only latent information) at each horizon (days ahead of prediction). Each prediction is an average of prediction using different lags (i.e. lags = 1, 3, 5, and 7).

Magnitude indicates a singular conclusion that it is possible to build a predictor for states of emotions that observe only physical actions and daily routines. This observable-only predictor not only yields good prediction results on its own, but its
prediction performance (in terms of magnitudes of error) is also even better than
the forecaster trained with both observable and latent information. However, some
of the observable information used in this thesis can still be cumbersome to cap-
ture in an automated manner (note that all observable information in the dataset
were collected by having the participants manually respond to the questionnaires.).
These observable attributes include number of sleep hours and number of drinks.
In the next immediate step, this observable information would still be input to
the forecaster manually either by human observers or the test subjects themselves,
while we continue to investigate the possibility to obtain this observable informa-
tion in an automated, sensor-based manner.

4.7.2 Impact of Different Lags

Table 4.4: Comparison of the forecasting performance between Random Forest
forecasters trained with data of different lags (i.e. $L = 1, 3, 5, 7$) on the select
latent attributes. The predictions of each horizon (i.e. 0 to 7) are averaged.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Mean Absolute Error (MAE)</th>
<th>Directional Accuracy (DAC)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L1</td>
<td>L3</td>
</tr>
<tr>
<td>CONTROL_d</td>
<td>8.04343</td>
<td>8.10174</td>
</tr>
<tr>
<td>NegAffect_d</td>
<td>6.05397</td>
<td>6.07227</td>
</tr>
<tr>
<td>PHYHEALTH_d</td>
<td>8.80460</td>
<td>8.80975</td>
</tr>
<tr>
<td>PosAffect_d</td>
<td>6.07272</td>
<td>6.07991</td>
</tr>
<tr>
<td>SATHEALTH_d</td>
<td>8.23313</td>
<td>8.23469</td>
</tr>
<tr>
<td>SATLIFE_d</td>
<td>7.42245</td>
<td>7.43554</td>
</tr>
</tbody>
</table>

It is often a natural question when implementing a time series forecaster: how
much history of data does the forecaster need to take into account to achieve the
optimal forecasting performance? In this section, the observable-only forecaster,
which implements the Random Forest algorithm, is trained with observable infor-
mation with different lags of 1, 3, 5, and 7 days. Mean absolute error and directional
accuracy are measured from the average forecasting results of the horizons 0, 1, 2,
..., 7. Table 4.4 lists the results for all the select six attributes.

In terms of mean absolute error, the lag of one day yields the best performance
on CONTROL_d and SATLIFE_d, the lag of five days on PHYHEALTH_d,
PosAffect_d, and SATHEALTH_d, and the lag of seven days on NegAffect_d.
Even though there is no single dominant consensus on the best lag that would
yield the minimum error for all the attributes, the results illustrate that differ-
ent attributes are dependent on different lengths of past data. For example,
the most recent information is already good enough to determine the states of
CONTROL_d and SATLIFE_d, which represent the level of perceived control
over surrounding environments and the level of satisfaction in life, respectively.
While, more information is needed in the case of PHYHEALTH_d, PosAffect_d,
and $SATHEALTH_d$. In some cases where the attributes tend to exhibit seasonality (e.g., $NegAffect_d$) a lag of one week (7 days) would be the optimal. This particular case of $NegAffect_d$ is quite intuitive since an individual tend to experience depression-induced events (meetings, homework dues, etc.) in a weekly basis.

It is, however, interesting to note that the lag of one day yields a dominant result in terms of directional accuracy. This suggests that the relational emotional states in the next time period can be best predicted using the information in the most recent time period. The Random Forest forecaster yields the directional accuracy of roughly 68% on average. This means that, it can predict whether a particular emotional attribute will be higher or lower than the current day with the accuracy of 68%. Though this number is not high enough to implement in the real system, this is a primary investigation that leaves an ample room for improvement.
Chapter 5

Conclusion and Future Direction

The research reported in this thesis illustrates a preliminary effort to use machine learning techniques to analyze a set of psychological multivariate time series data. The problem is framed as a multivariate time series forecasting problem, in which multiple forecasting models can be compared using standard forecasting evaluation protocols. The selected models comprise the traditional vector autoregression (VARX) and multiple variants of machine learning based regression models that are modified for multivariate time series data. The models are compared on the ability to forecast the values of the selected six latent attributes, using a set of multi-variate time series psychological data collected from 150 test participants in State College, PA. Such a dataset surveys each participant’s daily emotional, physical, and behavioral states. The first set of experiments shows that the Random Forest based forecaster is best suited for the dataset. In the second batch of experiments, it is empirically shown that some of the latent attributes could be effectively predicted using only the observable information. These results not only provide primary, promising initial progress towards our ultimate goal, but also point out to ample of room for improvement. Future works could include collecting more data, and investigating automatic methods to collect participants’ information (e.g. using Microsoft Kinect\(^1\) to detect movements and Apple Health app\(^2\) to detect sleep pattern and other physical information).

\(^1\)https://en.wikipedia.org/wiki/Kinect  
Bibliography


Vita

Suppawong Tuarob

Suppawong Tuarob was born in Trang Province of Thailand on December 25th, 1986. After completing his high school at Mahidol Wittayanusorn school in Nakorn Prathom, Thailand, he attended another year of high school at Westtown School, West Chester, PA. He completed both his Bachelor and Master of Engineering in Computer Science and Engineering from the University of Michigan-Ann Arbor in December 2009 and May 2010 respectively. He joined the PhD program in Computer Science and Engineering at the Pennsylvania State University, and passed his dissertation defense in October 2014. Thereafter, he joined the MS program in Industrial and Manufacturing Engineering at the Pennsylvania State University to expand his research horizon in October 2014. He became a member of the Design Analysis Technology Advancement (D.A.T.A.) Laboratory supervised by Dr. Conrad Tucker. His research interest lies in information retrieval, applied machine learning, data (especially text) mining, statistical topic modeling, social media, sentiment analysis, product design informatics, and time series analysis.

Selected Representative Publications


