J Healthc Inform Res
https://doi.org/10.1007/s41666-018-0017-x

RESEARCH ARTICLE

# Switching-State Dynamical Modeling of Daily Behavioral Data

**Randy Ardywibowo[1] · Shuai Huang[2] · Shupeng Gui[3] · Cao Xiao[4] · Yu Cheng[4] · Ji Liu[3] · Xiaoning Qian[1]**

Received: 31 August 2017 / Revised: 29 January 2018 / Accepted: 27 February 2018


**Abstract** Emerging wearable and environmental sensor technologies provide health professionals with unprecedented capacity to continuously collect human behavioral data for health monitoring and management. This enables new solutions to mitigate globally emerging health problems such as obesity. With such outburst of dynamic sensor data, it is critical that appropriate mathematical models and computational methods are developed to translate the collected data into accurate characterization of the underlying health dynamics, enabling more reliable personalized monitoring,

✉ Randy Ardywibowo
randyardywibowo@tamu.edu

Shuai Huang
shuaih@uw.edu

Shupeng Gui
shupenggui@gmail.com

Cao Xiao
cxiao@us.ibm.com

Yu Cheng
chengyu@us.ibm.com

Ji Liu
ji.liu.uwisc@gmail.com

Xiaoning Qian
xqian@ece.tamu.edu

[1] Texas A&M University, College Station, TX 77840, USA

[2] University of Washington, Seattle, WA 98195, USA

[3] University of Rochester, Rochester, NY 14620, USA

[4] IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA



Published online: 29 March 2018

prediction, and intervention of health status changes. In addition to addressing common analytic challenges in analyzing sensor behavioral data, such as missing values and outliers, we focus on modeling heterogeneous dynamics to better capture health status changes under different conditions, which may lead to more effective state-dependent intervention strategies. We implement switching-state dynamic system models with different complexity levels on real-world daily behavioral data. Evaluation experiments of these models are conducted to demonstrate the importance of modeling the dynamic heterogeneity, as well as simultaneously conducting missing value imputation and outlier detection in achieving interpretable health dynamic models with better prediction of health status changes.

**Keywords** Switching-state dynamic systems · Daily behavioral data analysis · Mobile health · Longitudinal patient health modeling · Missing data and outlier treatment

## 1 Introduction

Currently, obesity is considered a public health issue as over one third of the US adult population is classified as obese [1]. However, addressing obesity is believed to be beyond the capacity of the healthcare industry [2], motivating the development of smart and scalable health solutions that can automate personalized activity planning.

Smart health solutions are becoming ever more feasible with the rapid development of sensors and mobile applications that can continuously collect human behavioral data such as physical activity, food intake, and body mass index (BMI) [3]. However, with such outburst of dynamic sensor data, several challenges arise in translating them into personalized health monitoring and activity plans effectively. Besides common challenges in analyzing sensor behavioral data, such as missing values and outliers, modeling the complex health dynamics with potential influence from human daily behaviors also poses significant challenges.

We implement a switching-state auto-regressive (SAR) population model [4] to capture the complex interactions of human daily behaviors. We have adopted this model framework due to its capability to capture instantaneous changes in human activity and to classify inherent health stages in a population. We compare our model to a similar dynamic model that does not consider these factors, showing that considering the switching-state behavior and population-wide effects improves the model's prediction performance significantly.

To handle missing values and outliers in daily behavioral data, we simultaneously consider missing value imputation and outlier detection while conducting model identification. We compare our simultaneous imputation and outlier detection method with typical data preprocessing approaches, showing that integrating missing value imputation and outlier detection with model identification significantly improves model accuracy. The preprocessing methods we compare include off-the-shelf missing value imputation and outlier detection methods, such as mean imputation and median filters, as well as analytic methods based on functional data analysis methods, such as functional principal component analysis (FPCA) [5, 6].

Finally, we conduct evaluation experiments to obtain the most parsimonious SAR model with the learned model parameters based on a real-world daily behavioral data

set, which shows improved prediction accuracy of BMI temporal changes with different daily activity profiles.

## 2 Methods

In our presentation, we adopt the following notation: regular lowercase letters denote scalars and boldface lowercase letters represent column vectors. When not explicitly specified, subscripts index time while superscripts index subjects. We use a colon (:) when we refer to a group of variables at different indices. For example, $x_{1:t}^{1:n}$ refers to the variable $x$ at time 1 to $t$ of subjects 1 through $n$.

### 2.1 Switching-State Auto-Regressive Population Model

We implement a population switching-state auto-regressive model in our analysis of the daily behavioral dataset. To model the potential heterogeneous dynamic changes of health status for each subject under study, we assume that the underlying dynamic system can switch between different dynamic behaviors under different conditions at different times. For the $i^{\text{th}}$ subject ($i \in \{1, \ldots, N\}$) at time $t$, we assume that there exists a discrete latent health state $s_t^i$ determining the dynamics of a health indicator, represented by $x_t^i$, which is also influenced by $p$ input variables capturing daily life behavior, denoted by $\boldsymbol{\upsilon}_t^i = \left[\upsilon_{t,1}^i, \upsilon_{t,2}^i, \ldots, \upsilon_{t,p}^i\right]^{\mathrm{T}}$. Specifically, in this paper, we are interested in the observed health indicator BMI as the health status of interest, and its change across time. The input variables include subjects' daily behavioral data, such as calorie intake (food), calories burned during workout or exercise, and workout time.

The SAR model is an extension of the classical auto-regressive (AR) model, which describes the time evolution of a variable that depends linearly on its past realizations, defined as follows:

$$x_t^i = a^{\mathrm{T}} x_{t-1}^i + \mathbf{b}^{\mathrm{T}} \boldsymbol{\upsilon}_t^i + c + \eta_t^i \tag{1a}$$

$$\eta_t^i \sim \mathcal{N}\left(0, \sigma_i^2\right) \tag{1b}$$

Here, $a$, $\mathbf{b}$, $c$, and $\sigma_i^2$ are the system coefficients of the AR model and the white noise variance respectively. Extending the above formulation, the SAR model allows these model parameters to be determined by a latent state $s$, denoted explicitly as $a(s)$, $\mathbf{b}(s)$, $c(s)$, and $\sigma_i^2(s)$. Specifically, SAR models the BMI dynamics by the following system model:

$$x_t^i = a\left(s_t^i\right)^{\mathrm{T}} x_{t-1}^i + \mathbf{b}\left(s_t^i\right)^{\mathrm{T}} \boldsymbol{\upsilon}_t^i + c\left(s_t^i\right) + \eta_t^i \tag{2a}$$

$$\eta_t^i \sim \mathcal{N}\left(0, \sigma_i^2\left(s_t^i\right)\right) \tag{2b}$$

In general, the system in (2) can incorporate any order of time lags $L_x$ and $L_u$ to model the potential high-order dependence relationships so that the model can be extended as follows:

$$x_t^i = \mathbf{a}\left(s_t^i\right)^{\mathrm{T}}\mathbf{x}_{t-1}^i + \mathbf{b}\left(s_t^i\right)^{\mathrm{T}}\mathbf{u}_t^i + c\left(s_t^i\right) + \eta_t^i \tag{3a}$$

$$\mathbf{x}_{t-1}^i = \left[x_{t-1}^i, x_{t-2}^i, \ldots, x_{t-L_x}^i\right]^{\mathrm{T}} \tag{3b}$$

$$\mathbf{u}_t^i = \left[\left(\upsilon_t^i\right)^{\mathrm{T}}, \left(\upsilon_{t-1}^i\right)^{\mathrm{T}}, \ldots, \left(\upsilon_{t-L_u+1}^i\right)^{\mathrm{T}}\right]^{\mathrm{T}} \tag{3c}$$

In this paper, we adopt a *population SAR model* assuming that the system coefficients $\mathbf{a}(s)$, $\mathbf{b}(s)$, and $c(s)$ are shared between subjects while each subject has independent measurement noise variance $\sigma_i^2(s)$. This treatment of measurement noise is reasonable, as each subject may have different levels of fluctuation in their daily behavior changes. Additionally, the subjects may also log their daily behaviors differently, with varying degrees of noise intensity.

For the case with $L_x = L_u = 1$, the population SAR model is illustrated in Fig. 1:

It has a finite Markov chain layer to model the health state changes along time and an AR model layer to capture the "controlled" dynamic changes at different health states, both assumed to be shared in the population under study. Clearly, introducing the hidden layer increases the model flexibility to enable the potential of modeling abrupt changes in human health status as well as daily activity. On the other hand, instead of assuming different subjects have their own independent dynamic models, the population model assumption controls the model complexity to avoid overfitting with the observed measurements and borrows signal strengths across subjects, especially considering the potential missing values and outliers in daily behavioral data.

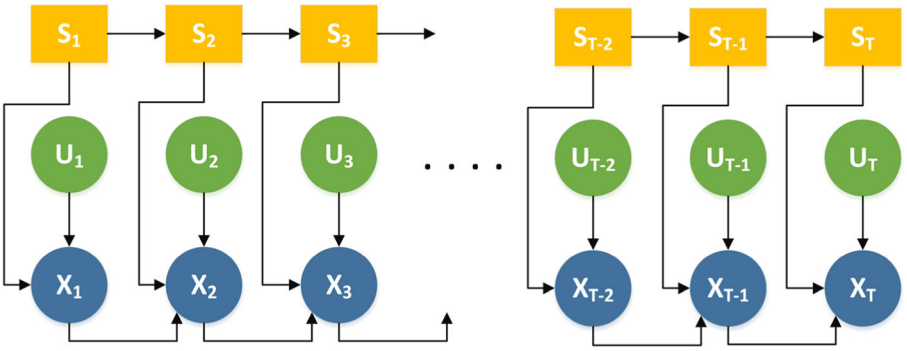


**Fig. 1** First-order SAR model

### 2.1.1 Learning the SAR Model

To learn such a population SAR model given the observed daily behavioral data and BMI changes, we have the following auto-regressive coefficients as well as health states to identify:

$$\theta = \left\{\mathbf{a}(s), \mathbf{b}(s), c(s), \sigma_i^2(s), s\in\{1,\ldots,\mathcal{S}\}, i\in\{1,\ldots,N\}\right\}. \tag{4}$$

As each subject's time series measurements are independent of each other given the population SAR model, we have the following likelihood function of the population SAR model given observed data:

$$p\left(X^{1:N}, S^{1:N}|U^{1:N},\theta\right) = \prod_{i=1}^{N} p_i\left(X^i, S^i|U^i,\theta\right) \tag{5a}$$

$$p_i\left(X^i, S^i|U^i,\theta\right) = p\left(x_1^i|\mathbf{u}_1^i, s_1^i,\theta\right)p\left(s_1^i\right) \times \prod_{t=2}^{T_i} p\left(x_t^i|\mathbf{x}_{t-1}^i, \mathbf{u}_t^i, s_t^i,\theta\right)p\left(s_t^i|s_{t-1}^i\right) \tag{5b}$$

Here, $X^i$, $U^i$, and $S^i$ are the health indicator (BMI), input covariates, and latent state values of subject $i$ at all time-points, while $T_i$ is the last time index of subject $i$.

To derive the maximum likelihood estimates (MLE) for model identification, expectation-maximization (EM) [4] is often adopted to find the set of system coefficients and variances $\mathbf{a}(s)$, $\mathbf{b}(s)$, $c(s)$, and $\sigma_i^2(s)$ for all $s \in \{1,\ldots,S\}$. This method alternates between estimating the state conditional probabilities $p\left(s_t^i|X^i, U^i\right)$ and optimizing the system coefficients based on the estimated state probabilities in the expectation and maximization steps respectively.

a) *E-step*: The expectation step is done by the forward-backward algorithm [4], which estimates the state probability $p\left(s_t^i|X^i, U^i\right)$ by combining partial solutions conditioned on past and future observations with respect to $t$. The partial solutions conditioned on past observations are denoted by $\alpha\left(s_t^i\right) = p\left(s_t^i, x_{1:t}^i\middle|\boldsymbol{u}_{1:t}^i,\theta\right)$, while the partial solutions for future observations are denoted by $\beta\left(s_{t-1}^i\right) = p\left(x_{t:T_i}^i|\boldsymbol{x}_{t-1}^i, \boldsymbol{u}_{t:T_i}^i, s_{t-1}^i,\theta\right)$. Given the model, we denote $p\left(x_t^i|\boldsymbol{x}_{t-1}^i, \boldsymbol{u}_t^i, s_t^i,\theta\right)$ by $\hat{p}_i\left(x_t^i|s_t^i\right)$:

$$\hat{p}_i\left(x_t^i|s_t^i\right) = \mathcal{N}\left(\mathbf{a}\left(s_t^i\right)^{\mathrm{T}}\mathbf{x}_{t-1}^i + \mathbf{b}\left(s_t^i\right)^{\mathrm{T}}\mathbf{u}_t^i + c\left(s_t^i\right), \sigma_i^2\left(s_t^i\right)\right) \tag{6}$$

Define the log-likelihood $L(\theta) = \log\left(\sum_i \sum_t \alpha\left(s_t^i\right)\right)$ with $\alpha\left(s_1^i\right) = p\left(x_1^i|\mathbf{u}_1^i, s_1^i,\theta\right) p\left(s_1^i\right)$. This can be efficiently solved as a filtering problem by the $\alpha$-recursion [7]:

$$\alpha\left(s_t^i\right) = \hat{p}_i\left(x_t^i|s_t^i\right) \sum_{s_{t-1}^i} p\left(s_t^i|s_{t-1}^i\right)\alpha\left(s_{t-1}^i\right). \tag{7}$$

On the other hand, the partial solution conditioned on future observations can be solved using the $\beta$-recursion:

$$\beta\left(s_{t-1}^{i}\right)=\sum_{s_{t}^{i}}\hat{p}_{i}\left(x_{t}^{i}|s_{t}^{i}\right)p\left(s_{t}^{i}|s_{t-1}^{i}\right)\beta\left(s_{t}^{i}\right), \tag{8}$$

and $\beta\left(s_{T_i}^{i}\right)=1$. By Bayes' rule, combining these two partial results yields the desired state probability:

$$\gamma\left(s_{t}^{i}\right)=p\left(s_{t}^{i}|X^{i},U^{i},\theta\right)=\frac{\alpha\left(s_{t}^{i}\right)\beta\left(s_{t}^{i}\right)}{\sum_{s_{t}^{i}}\alpha\left(s_{t}^{i}\right)\beta\left(s_{t}^{i}\right)}. \tag{9}$$

Because each subject's time series is conditionally independent with one another given the model, the expectation step can be done independently on each subject. Finally, we can derive the joint state transition probability for the hidden Markov chain layer by normalization with

$$p\left(s_{t}^{i},s_{t+1}^{i}|X^{i},U^{i},\theta\right)\propto\alpha\left(s_{t}^{i}\right)\hat{p}_{i}\left(x_{t+1}^{i}|s_{t+1}^{i}\right)\times p\left(s_{t+1}^{i}|s_{t}^{i}\right)\beta\left(s_{t+1}^{i}\right) \tag{10}$$

b) *M-step:* The maximization step uses the state distributions calculated in the expectation step to optimize the system coefficients by maximizing the likelihood:

$$E=\sum_{i}\sum_{t}\left\langle\log\hat{p}_{i}\left(x_{t}^{i}|s_{t}^{i}\right)\right\rangle_{p^{old}\left(s_{t}^{i}|X^{i},U^{i}\right)}+\sum_{i}\sum_{t}\left\langle\log p\left(s_{t}^{i}|s_{t-1}^{i}\right)\right\rangle_{p^{old}\left(s_{t}^{i},s_{t-1}^{i}\right)} \tag{11}$$

Rewrite the system coefficients and variables as follows:

$$\mathbf{d}\left(s_{t}^{i}\right)=\begin{bmatrix}\mathbf{a}\left(s_{t}^{i}\right)\\ \mathbf{b}\left(s_{t}^{i}\right)\\ c\left(s_{t}^{i}\right)\end{bmatrix}\qquad\mathbf{v}_{t-1}^{i}=\begin{bmatrix}\mathbf{x}_{t-1}^{i}\\ \mathbf{u}_{t}^{i}\end{bmatrix} \tag{12}$$

The Karush-Kuhn-Tucker (KKT) conditions [8] to maximize the likelihood with respect to $\mathbf{d}(s)$ lead to solving the following linear system by plugging (6) into (11):

$$\sum_{i}\sum_{t}p^{\text{old}}\left(s_{t}^{i}=\text{s}|X^{i},U^{i}\right)\frac{x_{t}^{i}\mathbf{v}_{t-1}^{i}}{\sigma_{i}^{2}(\text{s})}=\left[\sum_{i}\sum_{t}p^{\text{old}}\left(s_{t}^{i}=s|X^{i},U^{i}\right)\frac{\mathbf{v}_{t-1}^{i}\left(\mathbf{v}_{t-1}^{i}\right)^{\text{T}}}{\sigma_{i}^{2}(\text{s})}\right]\mathbf{d}(\text{s}) \tag{13}$$

Similarly, $\sigma_i^2$ may be solved by the following equation:

$$\sigma_{i}^{2}(\text{s})=\frac{1}{\sum_{t'}p^{\text{old}}\left(s_{t'}^{i}=\text{s}|X^{i},U^{i}\right)}\times\sum_{t}p^{\text{old}}\left(s_{t}^{i}=s|X^{i},U^{i}\right)\left[x_{t}^{i}-\mathbf{d}\left(s_{t}^{i}\right)^{\text{T}}\mathbf{v}_{t-1}^{i}\right]^{2} \tag{14}$$

### *2.1.2 Simultaneous System Identification, Missing Value Imputation, and Outlier Detection for SAR*

One of the critical challenges to learn the SAR model parameters arise from the large number of missing values and frequent outlier points in the data set. This is illustrated in a fragment of real-world time series BMI measurements in Fig. 2. Inappropriate handling of missing values and outliers may lead to computational difficulties from the holes in the data set, as well as the bias and loss of precision due to distortion of the data distribution [9]. For example, among the approaches that handle missing values [10], the mean imputation method ignores the context as it fails to utilize the underlying dynamics of the variables. The last-value-carried-forward method takes a conservative approach, underestimating the changes over time. Thus, neither of them is suitable for imputing missing values in the dynamic modeling context for human daily behavioral data.

Extending the SAR population model, we develop a method that can simultaneously remove outliers and impute missing values while conducting SAR model identification. We achieve this by modifying the maximization step of the previously introduced EM algorithm. For clarity of presentation, we firstly assume that there is only one input variable, so that $\boldsymbol{\upsilon}_t^i = \upsilon_{t,1}^i = \upsilon_t^i$. We will remove this restriction accordingly, as we shall see that in our method, each input variable can be handled separately.

The missing value imputation and outlier detection is formulated as follows: denote the state observations and input actions for subject $i$ as $X^i = \left[x_1^i, x_2^i, \ldots, x_{T_i}^i\right]$ and $U^i = \left[\upsilon_1^i, \upsilon_2^i, \ldots, \upsilon_{T_i}^i\right]$ respectively. Let $\Omega_{X_i}$ and $\Omega_{U_i}$ be the index set of observed elements of $X^i$ and $U^i$ respectively. We estimate $\hat{X}^i = \left[\hat{x}_1^i, \hat{x}_2^i, \ldots, \hat{x}_{T_i}^i\right]$ and $\hat{U}^i = \left[\hat{\upsilon}_1^i, \hat{\upsilon}_2^i, \ldots, \hat{\upsilon}_{T_i}^i\right]$ for system identification by solving the following optimization problem:

$$\min_{\hat{X},\hat{U}} \sum_{i=1}^{N} \sum_{t=2}^{T_i} \left\| x_t^i - \left[\mathbf{a}\left(s_t^i\right)^{\mathrm{T}} \mathbf{x}_{t-1}^i + \mathbf{b}\left(s_t^i\right)^{\mathrm{T}} \mathbf{u}_t^i + c\left(s_t^i\right)\right] \right\|^2 \tag{15a}$$

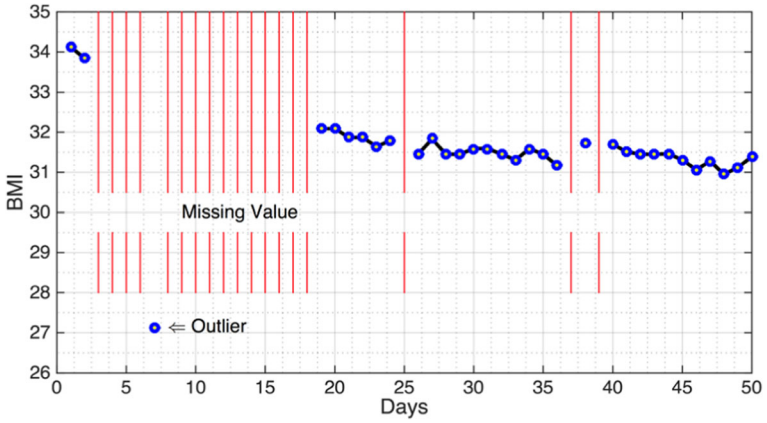


**Fig. 2** A typical example of life behavioral data from mobile sensors

$$\text{s.t.} \quad \left\| \left( \hat{X}^{i} - X^{i} \right)_{\Omega_{X_i}} \right\|_0 \leq \eta_X, \left\| \left( \hat{U}^{i} - U^{i} \right)_{\Omega_{U_i}} \right\|_0 \leq \eta_U \tag{15b}$$

The objective function (15a) is a squared loss function to evaluate the goodness-of-fit of the missing values and outlier estimates of the entire data set $\hat{X}$ and $\hat{U}$. Meanwhile, the constraints (15b) serve to limit the maximum number of outliers to be detected in $X$ and $U$. The values of $\eta_X$ and $\eta_U$ can be estimated by the upper bound of the percentage of outliers.

*2.1.3 Solution Strategy*

To simultaneously learn the system coefficients, estimate the state probabilities, as well as impute missing values and remove outliers, we alternatively optimize three groups of variables: the state distributions $p\left(s_t^i | X^i, U^i\right)$, the system coefficients $\theta = \left\{ \mathbf{d}(s),\ \sigma_i^2(s) \right\}$, and the missing value and outlier estimates $\left\{ \hat{X}^i, \hat{U}^i \right\}$ for all $i \in \{1, \ldots, N\}$ [11]. Calculating the state distributions and optimizing $\theta$ can be done based on the EM algorithm. On the other hand, the missing values and outliers for each subject $i$ are estimated using the projected gradient descent method as follows:

$$\hat{X}^{i}_{(k+1)} = \arg\min_{\hat{X}_i} \left\{ \left\| \hat{X}^{i} - \left( \hat{X}^{i}_{(k)} - \Delta g_{\hat{X}^{i}_{(k)}} \right) \right\|_F^2 \right\} \qquad \text{s.t.} \quad \left\| \left( \hat{X}^{i} - X^{i} \right)_{\Omega_{X_i}} \right\|_0 \leq \eta_X \tag{16}$$

Here, $g_{\hat{X}^{i}_{(k)}}$ is the partial derivative of the objective function with respect to $\hat{X}^{i}_{(k)}$, $\Delta$ is the step size that could be chosen to be a sufficiently small constant, while $\|\cdot\|_F$ denotes the Frobenius norm. The optimization procedure is done as follows: First, select $\eta_X$ elements in $\left( \hat{X}^{i}_{(k)} - \hat{X}^{i} - \Delta g_{\hat{X}^{i}_{(k)}} \right)_{\Omega_{X_i}}$ with the largest magnitudes as the outliers at the current iteration, forming a set $Z_{X_i}$. Second, assign the set of missing values $\bar{\Omega}_{X_i}$ and the set of detected outliers $Z_{X_i}$ with new estimates as $\left( \hat{X}_{(k+1)}{}^{i} \right)_{\bar{\Omega}_{X_i} \cup Z_{X_i}} = \left( \hat{X}^{i}_{(k)} - \Delta g_{\hat{X}^{i}_{(k)}} \right)_{\bar{\Omega}_{X_i} \cup Z_{X_i}}$. The remaining elements in $\hat{X}_{(k+1)}{}^{i}$ are set to the same values as $\hat{X}^{i}_{(k)}$. The update for $\hat{U}^{i}$ follows a similar procedure. Here, note that additional input variables can be separately handled by optimizing them with a similar procedure. The entire model identification, missing

value imputation, and outlier detection procedure is summarized in Algorithm 1.

**Algorithm 1** Population SAR Model Identification, Missing Value Imputation, and Outlier Detection

Input: $X^i_{\Omega_{X_i}}, U^i_{\Omega_{U_i}}, \eta_X, \eta_U \ \forall i \in \{1, \dots, N\}$

Output: $\mathbf{a}(s), \mathbf{b}(s), c(s), \sigma_i^2(\mathrm{s}), \hat{X}^i, \widehat{U}^i,$
$\forall i \in \{1, \dots, N\}, \forall s \in \{1, \dots, S\}$

Randomly initialize $\mathbf{a}(s)$, $\mathbf{b}(s)$, $c(s)$, $\sigma_i^2(\mathrm{s})$, and $p\left(s_t^i \middle| s_{t-1}^i\right) \ \forall i \in \{1, \dots, N\}$

Initialize $\left(\hat{X}^i\right)_{\overline{\Omega}_{X_i}}$ and $\left(\widehat{U}^i\right)_{\overline{\Omega}_{U_i}} \ \forall i \in \{1, \dots, N\}$ to the mean of $X^i_{\Omega_{X_i}}$ and $U^i_{\Omega_{U_i}}$ respectively.

$k \leftarrow 0$

While $\left\|L_{(k+1)}(\theta) - L_{(k)}(\theta)\right\| > \varepsilon$

E-step: Estimate $\gamma\left(s_t^i\right)$ by (9), and $p\left(s_t^i \middle| s_{t-1}^i\right)$ by (10) $\forall i \in \{1, \dots, N\}, \forall t \in \{1, \dots, T\}$.

M-step: Optimize $\mathbf{a}(s)$, $\mathbf{b}(s)$, $c(s)$ by (13), and $\sigma_i^2(\mathrm{s})$ by (14).

For $i \in \{1, \dots, N\}$

Optimize $\hat{X}^i$: Select top $\eta_X$ elements in $\left(\hat{X}^i - X^i - \Delta g_{\hat{X}^i}\right)_{\Omega_{X_i}}$ forming the index set $Z_{X_i}$.

$\left(\hat{X}^i\right)_{\overline{\Omega}_{X_i} \cup Z_{X_i}} \leftarrow \left(\hat{X}^i - \Delta g_{\hat{X}^i}\right)_{\overline{\Omega}_{X_i} \cup Z_{X_i}}$

Optimize $\widehat{U}^i$: Select top $\eta_U$ elements in $\left(\widehat{U}^i - U^i - \Delta g_{\widehat{U}^i}\right)_{\Omega_{U_i}}$ forming the index set $Z_{U_i}$.

$\left(\widehat{U}^i\right)_{\overline{\Omega}_{U_i} \cup Z_{U_i}} \leftarrow \left(\widehat{U}^i - \Delta g_{\widehat{U}^i}\right)_{\overline{\Omega}_{U_i} \cup Z_{U_i}}$

End

$k \leftarrow k + 1$

End

Return $\mathbf{a}(s)$, $\mathbf{b}(s)$, $c(s)$, $\sigma_i^2(\mathrm{s})$, $\hat{X}^i$, $\widehat{U}^i$, $\forall i \in \{1, \dots, N\}$, $\forall s \in \{1, \dots, S\}$

### 2.2 Functional Data Analysis-Based Imputation and Outlier Detection

We compare our simultaneous missing value imputation and outlier detection method with methods that initially preprocess these data defects, instead of solving them together with model identification as formulated in (15). Specifically, in addition to other off-the-shelf simple missing value imputation and outlier detection methods, we are interested in functional data analysis (FDA) methods [5] for studying time series data in our application.

### 2.2.1 Functional Principal Component Analysis

In the FPCA framework, given a vector of observations **y**, we estimate the underlying function $x$ by the penalized least squares smoothing method that is formulated to minimize the following loss function [5]:

$$\mathrm{PENSSE}_m(\mathbf{y}|\mathbf{c}) = (\mathbf{y}-\Phi\mathbf{c})^{\mathrm{T}}\mathrm{W}(\mathbf{y}-\Phi\mathbf{c}) \tag{17a}$$

$$+\lambda\,\mathrm{PEN}_m(x) \tag{17b}$$

Note that the underlying function $x$ is expressed in a different basis system as $\mathbf{x} = \Phi\mathbf{c}$, where $\Phi$ is the basis matrix and **c** contains the coefficients representing **x**, the realizations of the function $x$ in the basis system defined by $\Phi$. We can see that (17a) is a weighted least squares estimation. The roughness penalty term (17b) is added to enforce smoothness on the estimation of $x$, with $\lambda$ being a penalty coefficient and $\mathrm{PEN}_m(x)$ as the square integration of the $m^{\mathrm{th}}$ derivative, a measure of a function's roughness, defined as $\mathrm{PEN}_m(x) = \int [D^m x(s)]^2 ds$. The order of derivative penalized here is the second or fourth order derivative [5]. We can subsequently substitute $\Phi\mathbf{c}$ for $x$ and express this roughness penalty in matrix form as follows: $\mathrm{PEN}_m(x) = \mathbf{c}'\mathrm{R}\mathbf{c}$, in which $\mathrm{R} = \int D^m\phi(s)D^m\phi(s)'ds$.

The loss function (17) is convex and solving this model leads to closed-form solutions by the KKT conditions, similar as classic ordinary or weighted least squares problems [5]. The weighting for the smoothness penalty can be determined by cross validation, using the penalty parameter that produces the best estimation accuracy by the cross-validation testing.

The smoothness assumptions of the estimated behavioral data may change as different variables have varying degrees of smoothness, such as BMI vs. the number of calories burned in a day. To comprehensively evaluate the performance, two different basis systems were explored with this method, the B-spline bases and Haar wavelet bases. These basis systems were chosen due to their stark contrasts, with the B-spline basis offering the smoothest estimation while the Haar basis can capture abrupt changes in the data.

*a) B-spline basis:* The B-spline basis system represents functional data as a combination of piecewise spline functions of a certain degree $d$, with the corresponding polynomials approximating the function along with their derivatives up to $d-1$ are constrained to be equal at these breakpoints or knots. This produces a smooth representation of the behavioral data. To accommodate for abrupt changes that may happen in behavioral data, multiple knots may be placed in a single time point. The equation for a spline function is as follows. Let $B_k(t,\tau)$ be a piecewise polynomial function defined by the breakpoint sequence $\tau$, with $k$ being the number of the largest knot positioned less than or equal to $t$. Let $K$ be the total number of subintervals used. Then, the spline function $S(t)$ is defined as

$$S(t) = \sum_{1\le k\le K} c_k B_k(t,\tau) \tag{18}$$

When estimating user behavioral data, we found that the best estimation was estimated when 4th order spline is used while imposing a second derivative roughness penalty.

*b) Haar wavelet basis:* The Haar wavelet basis system is formed by a sequence of square-shaped functions [12]. Its mother wavelet $\psi(t)$ and scale function $\phi(t)$ are as follows:

$$\psi(t) = \begin{cases} 1 & 0 \le t < \frac{1}{2} \\ -1 & \frac{1}{2} \le t < 1 \\ 0 & \text{otherwise} \end{cases} \tag{19a}$$

$$\phi(t) = \begin{cases} 1 & 0 \le t < 1 \\ 0 & \text{otherwise} \end{cases} \tag{19b}$$

The mother wavelet and scale function represent the basis system by different dilations and translations $n$ and $k$ respectively, as defined by the following equation:

$$\psi_{n,k}(t) = 2^{n/2}\psi(2^n t - k) \tag{20}$$

This function is put into the basis matrix $\Phi$, with the columns being a basis formed by certain nonnegative integer $n$ and $0 \le k \le 2^K - 1$, with $K$ determining the number of approximating functions used. This basis system is utilized with no roughness penalty term, reducing the estimation problem into a weighted least squares problem.

Two examples comparing the missing value imputation and outlier detection performance of our simultaneous system identification, missing value imputation, and outlier detection (SSMO) formulation (15) against FPCA using the two different basis systems are shown in Fig. 3. On the top panel, we see that with abundant and smooth data, the FPCA-based methods perform similarly to ours. However, with sparsely observed data containing suspicious outliers, our method can better capture the overall trend of BMI changes, as shown on the bottom panel of Fig. 3.

### *2.2.2 Principal Component Analysis Through Conditional Expectation*

Designed for analyzing sparse data, the principal component analysis through conditional expectation (PACE) model is a non-parametric model that gives the best approximation of the functional data for an individual subject by a linear combination of $k$ functional curves by borrowing information from the entire collection of subjects. Formulated similarly to FPCA methods, PACE models the data for the $i^{\text{th}}$ subject $X^i(t)$ as noisy sampled points from a collection of trajectories. These trajectories are assumed to be independent realizations of a smooth random function, with unknown mean function $\mathrm{E}[X^i(t)] = \mu(t)$ and covariance function $\mathrm{cov}(X^i(s), X^i(t)) = G(s, t)$. The domain of $X^i(t)$ is bounded on a closed time interval $T$. Assuming an $L^2$ orthogonal expansion of $G$ exists in terms of eigenfunctions $\phi_k$ and eigenvalues $\lambda_k$ with $G(s,$

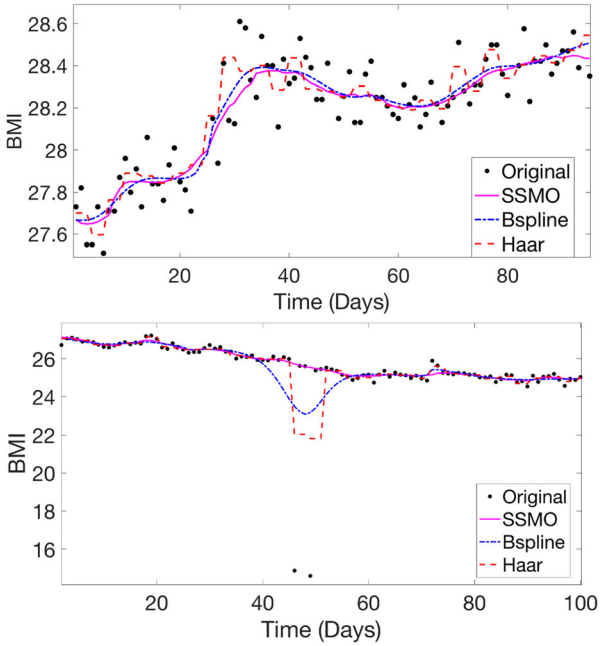


**Fig. 3** Missing value estimation and outlier detection using FPCA with B-spline and Haar wavelet basis

$t) = \sum_k \lambda_k \phi_k(s)\phi_k(t)$, the $i^{\text{th}}$ subject's trajectory can be represented as $X^i(t) = \mu(t) + \sum_k \xi_k^i \phi_k(t), t \in T$, where $\xi_k^i$ are uncorrelated random variables with zero mean. By also incorporating uncorrelated measurement errors, the PACE model can be formulated as follows:

$$Y_j^i = X^i\left(T_j^i\right) + \epsilon_j^i = \mu\left(T_j^i\right) + \sum_{k=1}^{\infty} \xi_k^i \phi_k\left(T_j^i\right) + \epsilon_j^i \;\; T_j^i \, \epsilon \, T, \tag{21}$$

where $\epsilon_j^i$ are uncorrelated measurement errors with mean zero and constant variance $\sigma_i^2$, and $Y_j^i$ is the $j^{\text{th}}$ observable data point of the $i^{\text{th}}$ subject.

To accommodate for the sparsity of daily behavioral data, local linear smoothers are used to estimate the mean function $\mu(t)$, instead of traditionally taking the average at each time point. This is because, in addition to being sparse, the time points of each user data may also not align with each other, causing bias in estimating the mean function through averaging. Estimation of the variance $\sigma_i^2$ is done through estimation of the covariance surface $\text{cov}\left(X\left(T_j^i\right), X\left(T_l^i\right)\right)$. A linear fit is used to estimate the diagonal elements of the covariance matrix, while a local quadratic fit is used for the off-diagonal elements, as the covariance matrix is maximal along its diagonal. The eigenfunctions can be subsequently found by discretizing the smoothed covariance surface. In these steps, we utilized the Gaussian kernel to perform the implicit feature mapping of the smooth surface estimation.

As a novelty introduced in the PACE formulation, the principal component scores are estimated by conditioning over the observations $\boldsymbol{Y}^i$, rather than through numeric

integration of the FPCA integral transform commonly used in traditional FPCA [13]. Compared to traditional FPCA, this is more suitable for sparse data since there are not enough points available to perform a numeric integration. This is estimated by the following equation:

$$\hat{\xi}_k^i = \hat{E}\left[\xi_k^i | \boldsymbol{Y}^i\right] = \hat{\lambda}_k \left(\hat{\phi}_k^i\right)^{\mathrm{T}} \hat{\Sigma}_{\boldsymbol{Y}^i}^{-1} \left(\boldsymbol{Y}^i - \hat{\boldsymbol{\mu}}^i\right) \tag{22}$$

Here, $\hat{\lambda}_k$, $\hat{\phi}_k^i$, $\hat{\boldsymbol{\mu}}^i$, and $\hat{\Sigma}_{Y^i}^{-1}$ are the estimates of $\lambda_k$, $\phi_k^i$, $\boldsymbol{\mu}^i$, and $\Sigma_{\boldsymbol{Y}^i}^{-1}$, the covariance matrix of $\boldsymbol{Y}^i$, respectively [6, 13]. We apply this formulation for each measured variable separately, estimating the model described above for each type of measured data in our collection of sensor behavioral data (calories burned, calories consumed, number of steps taken, workout time, and BMI). To select the number of eigenfunctions used in our model, we measure the fraction of variance explained (FVE) and pick the model that explains at least 95% of the total variation.

Figure 4 illustrates the comparison of imputed trajectories by SSMO and PACE for the same two subjects as in Fig. 3. Similar to the previous comparison with B-spline and wavelet based FPCA methods, our method performs comparably to PACE when we have abundant and smooth measurements. For cases with significant missing values and outliers, although PACE can be more robust compared to the previous FPCA methods, our SSMO method again captures the BMI changes more faithfully.

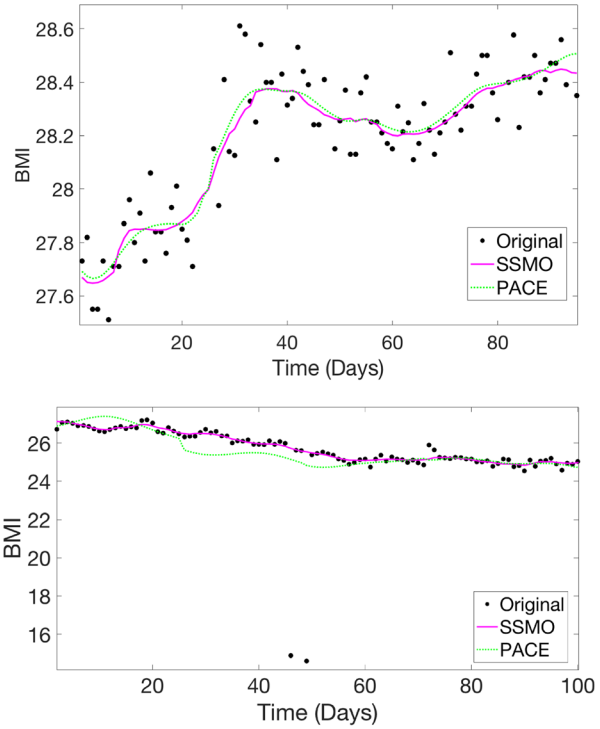


**Fig. 4** Missing value estimation and outlier detection using PACE

## 3 Results

We have implemented the population SAR model with the SSMO solution strategy on a real-world daily behavioral dataset that we collected. This dataset consists of daily fitness behaviors of 25 different users. The dataset includes diet, exercise information, and BMI collected from various sensor devices. In this dataset, almost all users show significant missing values and outliers, with patterns similar to Fig. 2. In our experiments, we take four types of recorded daily activity measurements, including calorie intake (food), calories burned during workout or exercise, and workout time.

In our evaluation experiments, we first illustrate that integrating missing value imputation and outlier detection with model identification outperforms the common two-step procedure of data preprocessing and model identification. We then evaluate the SAR models with different complexity levels and identify an appropriate model for the population dynamics in the given data set. We finally conduct a feature selection analysis to increase our model's robustness by removing potentially redundant covariates. We benchmark different models and methods by conducting one-step ahead prediction of future BMI trajectory.

### 3.1 Missing Value and Outlier Detection Evaluation

The simultaneous missing value and outlier detection of our method have been tested against several analytic and off-the-shelf imputation and outlier detection methods. The methods we have compared include the mean value imputation, last value-carried-forward imputation, functional principal component analysis (FPCA) with B-spline bases [5], and PACE for sparse data [6]. In addition, the commonly adopted simple mean value imputation and last value-carried-forward imputation methods are augmented with a median filter for outlier removal.

In our tests, the population SAR model with simultaneous imputation and outlier detection performs better than all the other methods we have benchmarked. As shown in Table 1, our population SAR model with the SSMO solution gives the best prediction accuracy overall. The functional data analysis methods PACE and FPCA with B-spline bases for two-step missing value imputation and outlier detection perform better than naive off-the-shelf methods, but our unified simultaneous imputation and outlier detection method in the population SAR model clearly outperforms them. A comparison of our method and the other functional data analysis methods benchmarked for imputation and outlier detection is shown in Fig. 5.

### 3.2 Model Selection

We evaluate several different model parameters on the order of time lags of the state observations, $L_x$, the order on the input variables, $L_u$, as well as the number of states in the SAR model $S$. We find that the parsimonious setup that gives the best accuracy is where $L_x = L_u = 1$, while $S = 3$. The prediction accuracy of different testing setups is shown in Table 2. Increasing $L_x$, $L_u$, or $S$ further did not yield any improvements in prediction accuracy. This "peaking" phenomenon in prediction accuracy may be caused by model overfitting of the training data. This causes the model to capture spurious dynamics with increased number of model parameters. This is undesirable, as

**Table 1** Evaluation result for different missing value estimation and outlier detection method

| | Mean + med | Last + med | Haar | B-spline | PACE | Simultaneous |
|---|---|---|---|---|---|---|
| RMSE | 0.5613 ± 0.5320 | 0.5294 ± 0.4875 | 0.1688 ± 0.1090 | 0.0826 ± 0.0621 | 0.0476 ± 0.0267 | *0.0321 ± 0.0166* |
| ABS | 0.2681 ± 0.1388 | 0.2588 ± 0.1346 | 0.0651 ± 0.0296 | 0.0443 ± 0.0229 | 0.0301 ± 0.0182 | *0.0241 ± 0.0116* |

The italicized entries indicate the best performing setup in terms of prediction accuracy

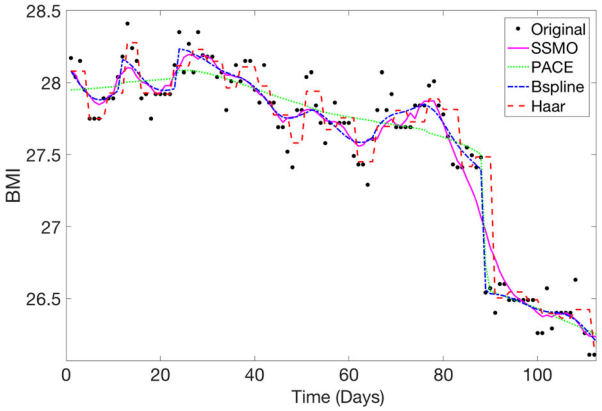


**Fig. 5** Missing value estimation and outlier detection method comparison

the additional states would only over fit the training data noise, and not capture the true health dynamics.

The corresponding system coefficients obtained for the parsimonious model with the best prediction accuracy are shown in Table 3. In Table 3, the coefficients for each variable are normalized such that each variable ranges from −1 to 1. With this normalization, the effects of each variable towards BMI can be directly seen without considering conversion factors.

For all three states, BMI would carry over to the next time point with very small changes, as the coefficients for BMI is close to one for all the states. In the identified model, the input variables that capture daily behavioral influence have less significant contribution to current BMI when compared to the effect of previous BMI. This makes intuitive sense: The inherent BMI change dynamics should be relatively stable, while the input variables should only produce incremental changes to the previous BMI.

**Table 2** Absolute one-step-ahead prediction error of the SAR population model under different model parameters (S1 denotes one-state model reducing to the traditional AR model, S2 denotes the model with two latent states, and S3 for the model with three latent states)

| | | $L_X$ | | |
|---|---|---|---|---|
| | | 1 | 2 | 3 |
| $L_U$ | 1 | S1: 0.045 ± 0.030 | S1: 0.072 ± 0.027 | S1: 0.084 ± 0.030 |
| | | S2: 0.029 ± 0.013 | S2: 0.037 ± 0.016 | S2: 0.052 ± 0.024 |
| | | *S3: 0.024 ± 0.012* | S3: 0.059 ± 0.056 | S3: 0.074 ± 0.072 |
| | 2 | S1: 0.049 ± 0.024 | S1: 0.059 ± 0.034 | S1: 0.084 ± 0.032 |
| | | S2: 0.029 ± 0.012 | S2: 0.040 ± 0.019 | S2: 0.064 ± 0.042 |
| | | S3: 0.031 ± 0.013 | S3: 0.041 ± 0.021 | S3: 0.051 ± 0.033 |
| | 3 | S1: 0.056 ± 0.023 | S1: 0.068 ± 0.026 | S1: 0.096 ± 0.060 |
| | | S2: 0.041 ± 0.015 | S2: 0.040 ± 0.017 | S2: 0.064 ± 0.044 |
| | | S3: 0.037 ± 0.026 | S3: 0.074 ± 0.072 | S3: 0.045 ± 0.030 |

The italicized entries indicate the best performing setup in terms of prediction accuracy

**Table 3** Normalized SAR coefficients for different variables under different states

| Variable | State 1 | State 2 | State 3 |
|---|---|---|---|
| BMI | 1.0003 | 0.9824 | 0.9950 |
| Exercise calories | − 0.0032 | − 0.0047 | − 0.0043 |
| Food calories | 0.0007 | 0.0187 | 0.0104 |
| Workout calories | 0.0031 | − 0.0080 | 0.0017 |
| Workout time | − 0.0251 | 0.0261 | 0.0072 |

We conjecture that state 2 represents the most active state; state 1 represents the least active state, while state 3 is an intermediary state in between these two. We speculate this due to the following observations: First, note that the coefficient for BMI for state 2, denoted by $a(2)$, is the smallest followed by $a(3)$ and $a(1)$. Furthermore, $a(2) < 1$ and $a(1) > 1$. This means that without any external intervention as observable input variables, subjects in state 2 inherently lose weight the fastest while subjects in state 1 inherently gain weight. Second, we observe that with increasing workout time, subjects in state 2 may have increasing BMI, but subjects in state 1 have decreasing BMI. We speculate that subjects in state 2 are gaining muscle mass while subjects in state 1 can better control their weight with more workout time.

The remaining coefficients also make sense intuitively. For example, for all the states, consuming food increases BMI while exercise helps control BMI.

### 3.3 Covariate Selection Through Correlation Analysis

We further conduct a correlation analysis of daily activity variables to reduce the model complexity by removing potentially redundant and/or strongly correlated covariates. With this, we hope to increase our model's robustness and better interpret the learned dynamic models under different conditions (states). We first analyze the pairwise correlation of four covariates, as shown in Table 4. We order the covariates based on their aggregated correlation with the other covariates and then sequentially remove the covariate and learn the corresponding population SAR models with the remaining covariates. This is repeated until a large drop in prediction accuracy is seen.

In our case, we start by removing the highest correlated feature, workout calories. Then, we removed exercise calories and workout time separately. We finally stopped when workout calories, exercise calories, and workout time were all removed, causing a great decrease in prediction error. The prediction errors for all the population SAR models with the corresponding covariates are shown in Table 5.

**Table 4** Pairwise covariate correlation

| | Exercise calories | Food calories | Workout calories | Workout time |
|---|---|---|---|---|
| Exercise calories | 1 | 0.1806 | 0.4042 | 0.1996 |
| Food calories | 0.1806 | 1 | 0.1143 | 0.1374 |
| Workout calories | 0.4042 | 0.1143 | 1 | 0.7160 |
| Workout time | 0.1996 | 0.1374 | 0.7160 | 1 |

**Table 5** Prediction accuracies of feature removal steps

| Features removed | ABS | RMSE |
|---|---|---|
| None | 0.0241 ± 0.0116 | 0.0321 ± 0.0166 |
| Workout calories | 0.0311 ± 0.0107 | 0.0410 ± 0.0163 |
| Workout calories, exercise calories | 0.0325 ± 0.0220 | 0.0414 ± 0.0288 |
| Workout calories, workout time | 0.0375 ± 0.0369 | 0.0539 ± 0.0782 |
| Workout calories, workout time, exercise calories | 0.0637 ± 0.1110 | 0.1406 ± 0.4140 |

Based on the prediction accuracy, we finalize our population SAR model with two covariates: calorie intake (food) and workout time. Its corresponding system coefficients are shown in Table 6. Here, we see that there are three clear modes for the subject's health dynamics. Subjects in state 1 tend to gain BMI more easily while subjects in state 2 tend to lose BMI more easily, as shown by their larger and smaller coefficients respectively. Finally, subjects in state 3 are resistant to BMI change due to behavioral actions while having a steadily decreasing baseline BMI shown by the state's coefficient for BMI, denoted by $a(3) < 1$.

### 3.4 Prediction Accuracy Evaluation

We further compare our model with the linear dynamic system model without switching states, denoted as SSMO. Unlike our model, this model does not consider the potential heterogeneous dynamic changes in daily behavioral data and models each subject's dynamics with a different model instead of adopting a population model.

We benchmark the two models using both the L-1 norm absolute difference error (ABS) and the residual mean squared error (RMSE) in conducting one-step ahead prediction of future BMI trajectory. Our tests show that the SAR population model performs significantly better than SSMO as shown in Fig. 6 and Table 7. Clearly, our population SAR model captures the BMI changes more faithfully by allowing abrupt changes and borrowing signal strengths across subjects.

## 4 Conclusions and Future Work

We have implemented and carried out comprehensive evaluation of population switching-state auto-regressive (SAR) models together with missing value imputation and outlier detection on real-world daily behavioral data. Different from the existing

**Table 6** Normalized SAR coefficients for the final selected feature set

| Variable | State 1 | State 2 | State 3 |
|---|---|---|---|
| BMI | 1.0007 | 0.9946 | 0.9992 |
| Food calories | 0.0051 | 0.0017 | 0.0001 |
| Workout time | − 0.0098 | − 0.0209 | −0.0013 |

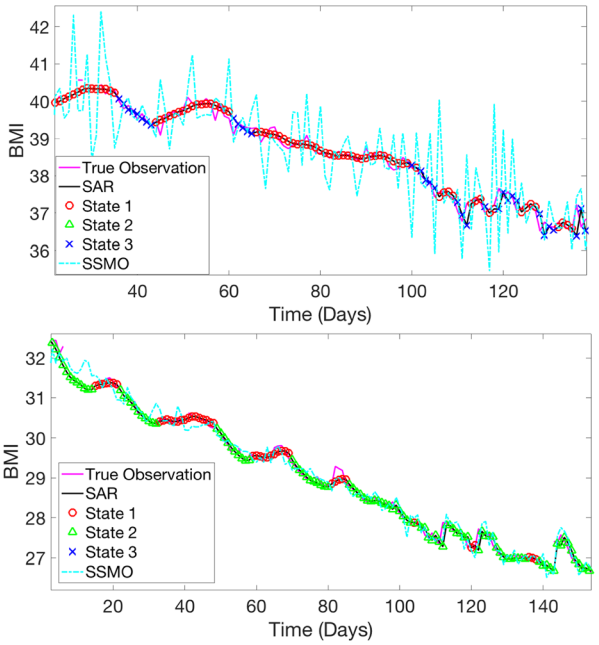


**Fig. 6** Prediction trajectory comparisons of the final SAR model

common procedure of imputation and outlier detection as separated data preprocessing step when analyzing behavioral sensor data, we handle missing data and outliers by simultaneously considering them while conducting model identification. We have conducted model selection to obtain the most accurate and parsimonious representation of the given data set and have shown that the identified model makes intuitive sense.

From our evaluation experiments, conducting missing value imputation and outlier detection while simultaneously identifying the model significantly improves model accuracy when compared with methods that firstly preprocess the data. In addition, we show that considering population-wide effects and dynamic heterogeneity significantly improves prediction performance on our data set.

As the dynamics of human behavioral data has been largely an uncharted research territory, characterizing the science of these unknown dynamics demands more in-depth study of the principles and complex relationships among the health outcomes and their control variables. By understanding these relationships, we plan to derive an automatic health intervention framework using the learned daily behavioral health model. Ultimately, integration of these highly analytic models in real-world clinical

**Table 7** Prediction accuracy comparison

| | SSMO | SAR |
|---|---|---|
| ABS | 0.2235 ± 0.2873 | *0.0325 ± 0.0220* |
| RMSE | 0.4032 ± 0.6140 | *0.414 ± 0.0288* |

The bold/italicized entries indicate the best performing setup in terms of prediction accuracy

implementation demands collaborations with systems engineering and health implementation science to ensure optimal patient treatment.

In addition to deriving personalized health management, the proposed system is generally useful for dynamic modeling with big and low-quality data and their translation into healthcare decision making outside of clinical settings. With appropriate infrastructure, it will have a profound impact on deriving effective smart and connected health solutions using emerging mobile sensors and applications.

**Funding Information** This work was partially supported by the National Science Foundation (NSF) grants Division of Communication and Computing Foundations (CCF) awards #1718513, #1715027, and #1714136. The authors thank them for their kind funding and support.

**Compliance with Ethical Standards**

**Conflict of Interest** The authors declare that they have no conflict of interest.

## References

1. Ogden CL, Carroll MD, Kit BK, Flegal KM (2014) Prevalence of childhood and adult obesity in the United States, 2011-2012. J Am Med Assoc 311:806–814
2. Rodgers GP, Collins FS (2012) The next generation of obesity research: no time to waste. J Am Med Assoc 308:1095–1096
3. Consolvo S, McDonald DW, Toscos T, Chen MY, Froehlich J, Harrison B et al (2008) Activity sensing in the wild: a field trial of ubifit garden. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, (pp. 1797-1806). ACM
4. Barber D (2012) Bayesian reasoning and machine learning. Cambridge University Press, New York
5. Ramsay JO (2006) Functional data analysis: John Wiley & Sons, Inc
6. Yao F, Müller H-G, Wang J-L (2005) Functional data analysis for sparse longitudinal data. J Am Stat Assoc 100:577–590
7. Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77:257–286
8. Boyd S, & Vandenberghe L (2004) Convex optimization. Cambridge University Press, New York
9. Somasundaram R, Nedunchezhian R (2011) Evaluation of three simple imputation methods for enhancing preprocessing of data with missing values. Int J Comput Appl 21(10):14–19
10. Gelman A, Hill J (2006) Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, New York
11. Tseng P (2001) Convergence of a block coordinate descent method for nondifferentiable minimization. J Optim Theory Appl 109:475–494
12. Debnath L, Shah FA (2002) Wavelet transforms and their applications. Birkhäuser, Boston
13. Mardia K, Kent J, Bibby J (1979) Multivariate analysis. Academic Press, London