Switching-state Dynamic Modeling of Daily Behavioral Data

UNDERGRADUATE RESEARCH SCHOLAR

Advisor: Dr. Xiaoning Qian‡

Randy Ardywibowo‡, Shuai Huang§, Cao Xiao◇, Shupeng Gui†, Yu Cheng◇, Ji Liu†

‡ Texas A&M University, § University of Washington, † University of Rochester, ◇ IBM T.J. Watson Research Center

Dwight Look College of ENGINEERING TEXAS A&M UNIVERSITY

Background

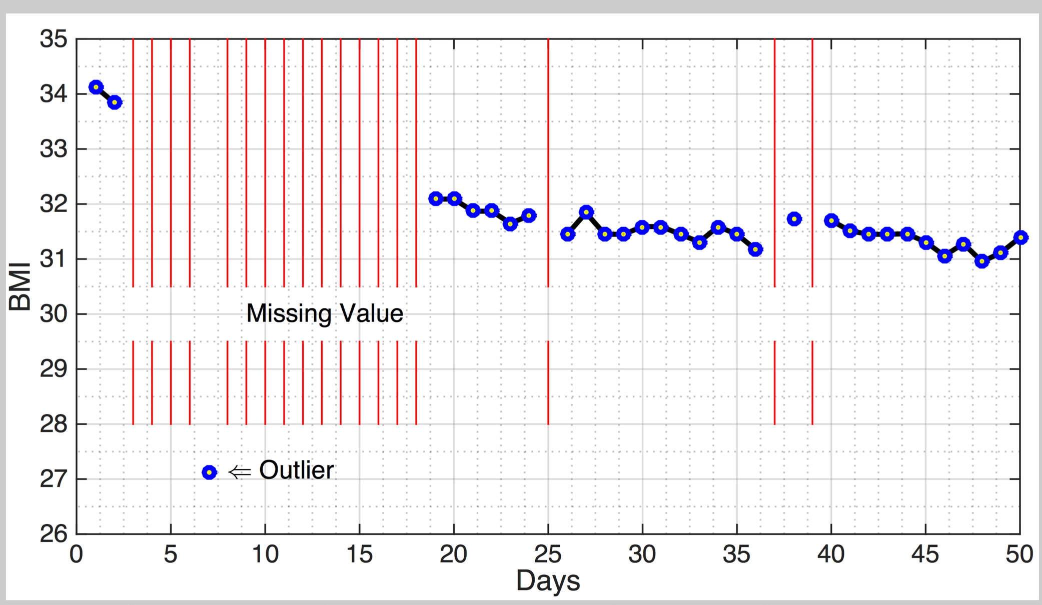
Personalized health monitoring and intervention may help mitigate global health problems such as obesity outside of clinic settings.

Sensors and mobile health Apps can monitor life behavior such as physical activity, food intake, body weight.

Example of a Health Monitoring App

Challenges

Besides common challenges in analyzing sensor behavior data, such as missing values and outliers, modeling the complex health dynamics influenced by human daily behaviors also pose significant challenges.



Outliers and Missing Values of BMI Data

Solution

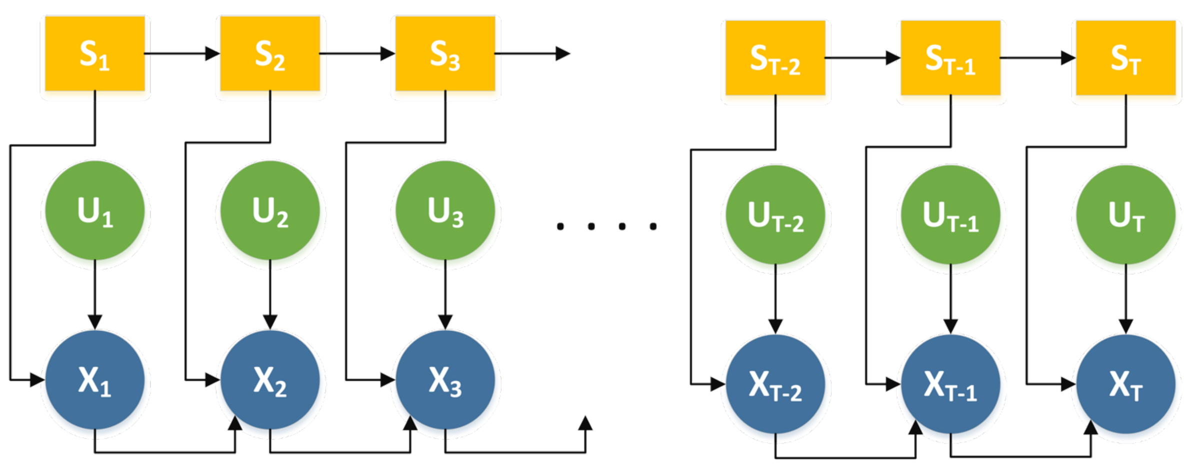
We implemented a Switching-state Auto-Regressive (SAR) population model to analyze daily behavioral data.

We implemented this model due to its capability to capture instantaneous changes in human activity and to classify inherent health stages in a population.

We tested this method against another dynamic system model that doesn't take the switching-state behavior and population-wide effects into account.

Population SAR Model

$$p_i(\vec{X}_i, \vec{S}_i | \vec{U}_i, \boldsymbol{\theta}) = \prod_{t=1}^{T} p(x_t | \hat{\mathbf{x}}_{t-1}, \hat{\mathbf{u}}_t, s_t, \boldsymbol{\theta}) p(s_t | s_{t-1})$$



Graphical representation of a 1st order SAR Model [1]

For the i^{th} subject at time t, We assume that there exists a discrete latent health state s_t^i determining the dynamics of a health indicator x_t^i, as well as the influence of input variables capturing daily life behavior, u_t^i.

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

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

Simultaneous Missing Value Imputation and Outlier Detection (SSMO)

We extend the SAR population model by developing a method that can remove outliers, impute missing values, while simultaneously conducting SAR model identification.

$$\min_{\hat{X}, \hat{U}} \sum_{i=1}^{N} \sum_{t=0}^{T-1} \left\| x_t^i - \hat{x}_t^i \right\|^2$$

$$\text{s.t.} \quad \left\| \left(\hat{X}_i - X_i\right)_{\Omega_{x_i}} \right\|_0 \leq \eta_x, \left\| \left(\hat{U}_i - U_i\right)_{\Omega_{u_i}} \right\|_0 \leq \eta_u$$

Where,

$$\hat{x}_t^i = \left[\left(\hat{\mathbf{x}}_{t-1}^i\right)^{\mathrm{T}} \mathbf{a}(s_t^i) + \left(\hat{\mathbf{u}}_t^i\right)^{\mathrm{T}} \mathbf{b}(s_t^i) + c(s_t^i) \right]$$

The optimization problem evaluates the goodness-of-fit of the imputation while the constraints serve to limit the maximum number of outliers in $\hat{X}$ and $\hat{U}$.

Solution Strategy

Expectation-Maximization (EM) is adopted to find the set of system coefficients and variances. This method recursively alternates between estimating the state conditional probabilities and optimizing the system coefficients. The E-step can be done by dynamic programming through the forward-backward algorithm [2].

The M-step optimizes the coefficients by minimizing the Kullback-Leibler (KL) divergence w.r.t. $\mathbf{d}(s) = [\mathbf{a}(s)^{\mathrm{T}}, \mathbf{b}(s)^{\mathrm{T}}, \mathrm{c}(s)^{\mathrm{T}}]^{\mathrm{T}}$ and $\sigma_i^2(s)$:

$$E = \sum_i \sum_t \langle \log p(x_t^i | \hat{\mathbf{x}}_{t-1}^i, \hat{\mathbf{u}}_t^i, \mathbf{d}(s_t^i)) \rangle_{p^{old}(s_t^i | x_{1:T}^i)} + \sum_i \sum_t \langle \log p(s_t^i | s_{t-1}^i) \rangle_{p^{old}(s_t^i, s_{t-1}^i)}$$

The missing values and outliers are optimized in the maximization step using the projected gradient descent method:

$$\hat{X}_i^{k+1} = \arg\min_{\hat{X}_i} \left\{ \left\| \hat{X}_i - \left(\hat{X}_i^k - \Delta g_{\hat{X}_i^k} \right) \right\|_F^2 \right\}$$

$$\text{s.t.} \quad \left\| \left(\hat{X}_i - X_i\right)_{\Omega_{x_i}} \right\|_0 \leq a$$

Here, $g_{\hat{X}_i^k}$ is the partial derivative of the objective function w.r.t. $\hat{X}_i^k$, Δ is the step size. We can optimize this by setting the outliers and missing values to their new estimates:

$$\left(\hat{X}_i^{k+1}\right)_{\overline{\Omega}_{x_i} \cup Z_{X_i}} = \left(\hat{X}_i^k - \Delta g_{\hat{X}_i^k}\right)_{\overline{\Omega}_{x_i} \cup Z_{X_i}}$$

The update for $\widehat{U}_i$ follows a similar procedure.

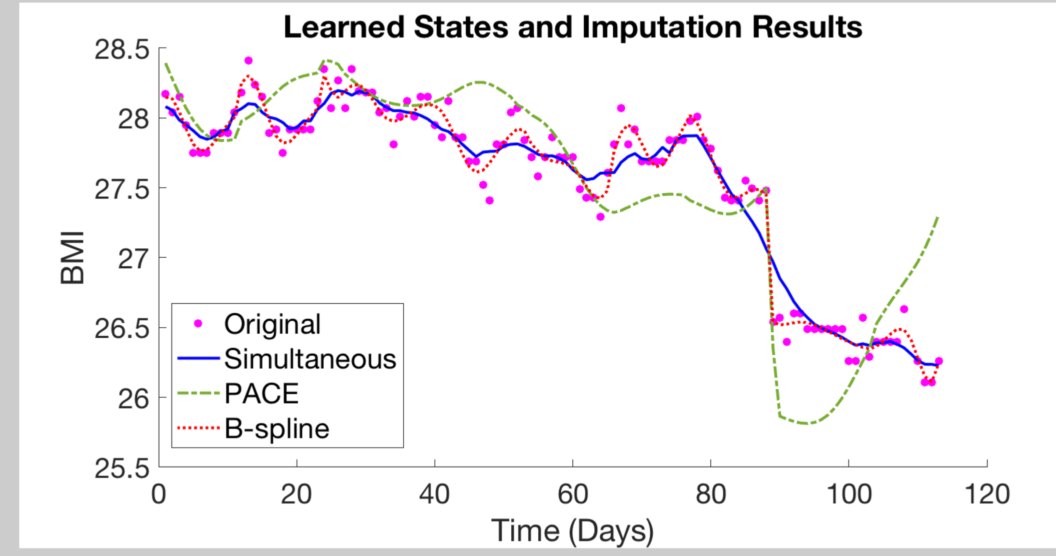
Model Selection

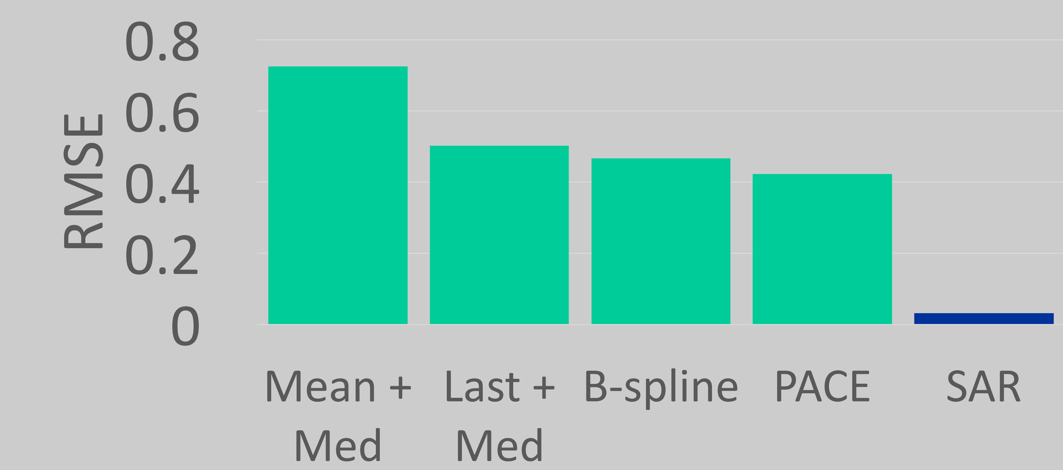
We compared several different sets of model parameters to obtain the most parsimonious setup that gives the best accuracy. This setup was found to be when $L_x = L_u = 1$, while $S = 3$.

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

Evaluation

The simultaneous imputation method was tested against off-the-shelf imputation methods as well as analytic imputation methods based on Functional Principal Component Analysis (FPCA) [4]. Our SSMO method was shown to be superior to all of the benchmarked methods.

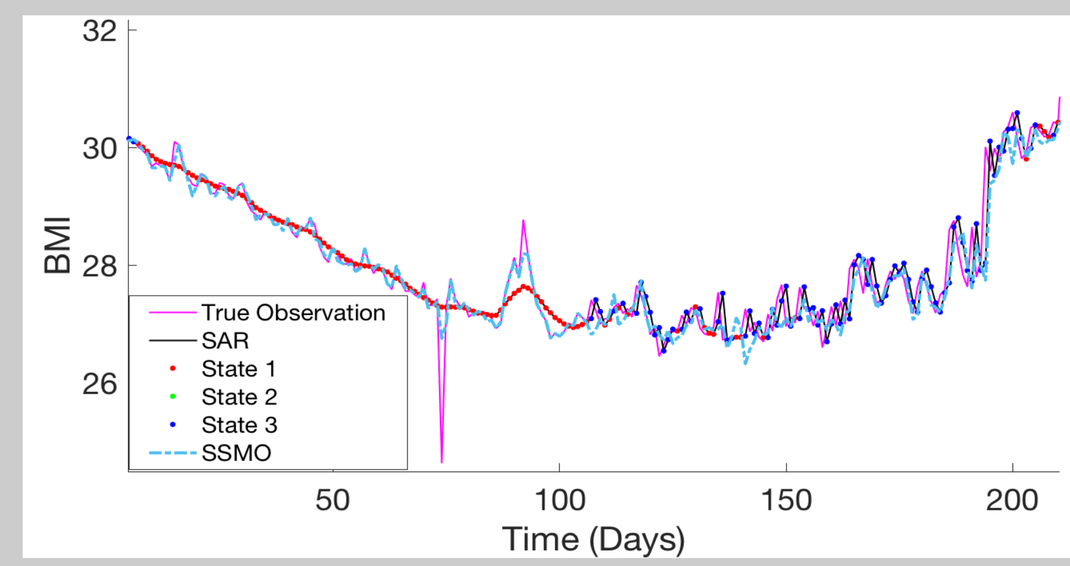




Prediction

The SAR population model was tested for prediction accuracy against a similar model that doesn't take the switching state behavior and population-wide effects into account [3]. Our tests showed that considering these factors significantly improved prediction accuracy.

	SSMO	SAR
ABS	0.22 ± 0.29	**0.024 ± 0.012**
RMSE	0.40 ± 0.61	**0.032 ± 0.017**



Predicted BMI

Conclusion

We presented a switching-state population model for daily behavioral data that can simultaneously impute missing values and detect outliers in the dataset. Our tests showed that these considerations significantly improved our model's prediction performance compared to existing methods.

[1] D. Barber, Bayesian reasoning and machine learning: Cambridge University Press, 2012.

[2] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proceedings of the IEEE, vol. 77, pp. 257-286, 1989.

[3] C. Xiao, S. Gui, Y. Cheng, X. Qian, J. Liu, and S. Huang, "Learning Longitudinal Planning for Personalized Health Management from Daily Behavioral Data", in submission, 2016.

[4] F. Yao, H.-G. Müller, and J.-L. Wang, "Functional data analysis for sparse longitudinal data," Journal of the American Statistical Association, vol. 100, pp. 577-590, 2005.